Yale University
EliScholar – A Digital Platform for Scholarly Publishing at Yale

Yale Graduate School of Arts and Sciences Dissertations

Spring 2021

Three Essays on the Role of Unstructured Data in Marketing Research

Ishita Sunity Kumar Chakraborty
Yale University Graduate School of Arts and Sciences, ishita.chakraborty@yale.edu

Follow this and additional works at: https://elischolar.library.yale.edu/gsas_dissertations

Recommended Citation
Chakraborty, Ishita Sunity Kumar, "Three Essays on the Role of Unstructured Data in Marketing Research" (2021). *Yale Graduate School of Arts and Sciences Dissertations*. 24.
https://elischolar.library.yale.edu/gsas_dissertations/24

This Dissertation is brought to you for free and open access by EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Yale Graduate School of Arts and Sciences Dissertations by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

Abstract

Three Essays on the Role of Unstructured Data in Marketing Research

Ishita Sunity Kumar Chakraborty

2021

This thesis studies the use of firm and user-generated unstructured data (e.g., text and videos) for improving market research combining advances in text, audio and video processing with traditional economic modeling. The first chapter is joint work with K. Sudhir and Minkyung Kim. It addresses two significant challenges in using online text reviews to obtain fine-grained attribute level sentiment ratings. First, we develop a deep learning convolutional-LSTM hybrid model to account for language structure, in contrast to methods that rely on word frequency. The convolutional layer accounts for the spatial structure (adjacent word groups or phrases) and LSTM accounts for the sequential structure of language (sentiment distributed and modified across non-adjacent phrases). Second, we address the problem of missing attributes in text in constructing attribute sentiment scores—as reviewers write only about a subset of attributes and remain silent on others. We develop a model-based imputation strategy using a structural model of heterogeneous rating behavior. Using Yelp restaurant review data, we show superior accuracy in converting text to numerical attribute sentiment scores with our model. The structural model finds three reviewer segments with different motivations: status seeking, altruism/want voice, and need to vent/praise. Interestingly, our results show that reviewers write to inform and vent/praise, but not based on attribute importance. Our heterogeneous model-based imputation performs better than other common imputations; and importantly leads to managerially significant corrections in restaurant attribute ratings.

The second essay, which is joint work with Aniko Oery and Joyee Deb is an information-theoretic model to study what causes selection in valence in user-generated

reviews. The propensity of consumers to engage in word-of-mouth (WOM) differs after good versus bad experiences, which can result in positive or negative selection of user-generated reviews. We show how the strength of brand image (dispersion of consumer beliefs about quality) and the informativeness of good and bad experiences impacts selection of WOM in equilibrium. WOM is costly: Early adopters talk only if they can affect the receiver's purchase. If the brand image is strong (consumer beliefs are homogeneous), only negative WOM can arise. With a weak brand image or heterogeneous beliefs, positive WOM can occur if positive experiences are sufficiently informative. Using data from Yelp.com, we show how strong brands (chain restaurants) systematically receive lower evaluations controlling for several restaurant and reviewer characteristics.

The third essay which is joint work with K.Sudhir and Khai Chiong studies success factors of persuasive sales pitches from a multi-modal video dataset of buyer-seller interactions. A successful sales pitch is an outcome of both the content of the message as well as style of delivery. Moreover, unlike one-way interactions like speeches, sales pitches are a two-way process and hence interactivity as well as matching the wavelength of the buyer are also critical to the success of the pitch. We extract four groups of features: content-related, style-related, interactivity and similarity in order to build a predictive model of sales pitch effectiveness.

Three Essays on the Role of Unstructured Data in Marketing Research

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Ishita Sunity Kumar Chakraborty

Dissertation Director: K. Sudhir

June 2021

Copyright © 2021 by Ishita Sunity Kumar Chakraborty

All rights reserved.

Acknowledgments

This thesis would not have been possible without the constant encouragement and support of my advisors, colleagues, friends and family.

I thank my advisors from the bottom of my heart for their generous time and support at every stage of my PhD journey. I especially thank my committee chair K. Sudhir for shaping my research agenda and constantly encouraging me to push my boundaries to take on new challenges. He helped me to identify my strengths and taught me that true courage lies in accepting and working on one's shortcomings. I thank Aniko Öry for being a great mentor, co-author and friend — I am lucky to have found all three in one person. I thank Dragomir Radev for introducing me to the wonderful world of Natural Language Processing. His NLP and Advanced NLP courses opened up my mind to the vast possibilities in which these methods can be applicable for social science research. I thank Kosuke Uetake for his calm and honest feedback at all stages of my research and especially during the job market. I have learnt a lot from each of one of you by engaging in stimulating and long discussions and sometimes simply by observing how you work.

I wholeheartedly thank all faculty members in Yale Marketing and especially the Quant Marketing group — Subrata Sen, Jiwoong Shin, Vineet Kumar, Soheil Ghili and Alexander Burnap for guiding me at every step of my transition from a naive PhD student to an academic researcher. Thanks for introducing me to your exceptional research and always being available to give me your valuable feedback on research as well as other aspects of academic life. The Marketing Seminar series taught by all of you has been the most valuable learning experience of my time at Yale SOM. Prof Subrata Sen was the first person to contact me about my acceptance at the program and till date remains invested in my progress and well-being. I am grateful to have found a mentor like him!

My learning would not have been even half as joyous and eventful had it not been for the wonderful cohort I had at SOM. While the pandemic does not allow us to meet

everyday, we remain connected in spirit and virtually. A big thank you to Hortense Fong, Minkyu Shin and Dai (Dan) Ruiting for being such great friends and co-researchers. We have been through a lot of ups and downs together but that makes us the best cohort! I am also extremely grateful to my fantastic co-authors Joyee Deb, Minkyung Kim, Atanu R. Sinha and Khai Chiong from whom I learnt a lot and continue to learn everyday. Minkyung is the best senior anyone could have asked for and had a deep influence on my research journey. I am also grateful to my extremely generous seniors Jungju Yu and Michelle Y. Lu for always being there for us.

Finally, I could not have thought of coming back to grad school without the unwavering support of my family, especially my late grandfather Mr. Harisadhan Mukherjee. This work is a manifestation of his dreams and aspirations; he will always remain my guiding angel. I thank my mom Sabarna Chakraborty and dad Dr. Sunity Chakraborty for instilling in me curiosity, discipline and most importantly work ethics from an early age and encouraging me daily to work on my research. I thank my husband Akshay for being my best friend, partner-in-crime and most fervent supporter as well as sharpest critic. He co-lived the life of a PhD student all these years and supported me in every way possible inspite of his busy work schedules. I thank my sister Nikita for being my greatest friend, confidante and inspiration. I thank my in-laws, all my cousins and extended family for their undying love and support. Today, I miss my loving aunt Alo Dutta the most, who we lost to Covid -19 and who has been a mother to me all my life. Everything will be incomplete without you!

To Ma and Baba.

Contents

1 Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Missing Attributes 1

1.1 Introduction 1

1.1.1 Challenges in Attribute Level Sentiment Scoring from Text 3

1.2 Related Literature 7

1.2.1 Text Analytics on UGC and Online Reviews 8

1.2.2 Missing Attributes (Attribute Silence) in Reviews 10

1.3 Converting Text into Numeric Attribute Sentiment Scores 12

1.3.1 Attribute Sentiment Classifier: The Lexicon Method 15

1.3.2 Attribute Sentiment Classifier: A Deep Learning Hybrid Convolutional-LSTM Model 18

1.3.3 Deep Learning Implementation: Important Choices 21

1.3.4 Performance Measures for Model Comparison 23

1.4 Analysis Of Structured Ratings Accounting For Missing Attributes 24

1.4.1 A Structural Model of Rating Behavior 25

1.4.2 Model Estimation and Missing Attribute Ratings 26

1.4.3 What Drives Attribute Silence? 28

1.5 Empirical Application 30

1.5.1 Data 30

1.5.2 Descriptive Evidence on Attribute Rating Behavior 33
1.6 Results 35
1.6.1 Attribute Sentiment Classification 35
1.6.2 Structural Model Estimates 39
1.6.3 Drivers of Attribute Silence (Missingness) 43
1.6.4 Validation of Imputation 47
1.6.5 Correction for Attribute Silence in Attribute Ratings 49
1.7 Conclusion 51

2 When do consumers talk? 53
2.1 Introduction 53
2.2 Literature Review 57
2.3 Model 59
2.4 Equilibrium Characterization 63
2.4.1 Word-of-Mouth subgame 63
2.4.2 Main results 67
2.5 Extensions 70
2.5.1 Dynamics 70
2.5.2 Private firm type 71
2.5.3 More than one follower 73
2.5.4 Idiosyncratic value 73
2.6 Empirical Evidence 74
2.6.1 Restaurant Industry and Review Platforms 75
2.6.2 Data Description and Summary Statistics 76
2.6.3 Supporting Evidence from Data 77
2.7 Conclusion 86

3 What makes sales pitches work? Using Multi-Modal Video Data to Uncover Success Factors **89**

3.1 Introduction 89

3.2 Related Literature 90

3.3 Background and Data 92

3.3.1 Buyer Types 94

3.3.2 Descriptive statistics 94

3.4 Empirical Framework 96

3.5 Results 100

3.6 Conclusion and Future Work 103

4 Appendix to Chapter 1 **105**

4.1 Mturk Experiment 105

4.1.1 LDA 106

4.2 A Hybrid CNN-LSTM Deep Learning Architecture 106

4.2.1 Additional Performance Metrics 115

5 Appendix to Chapter 2 **117**

5.1 Appendix: Data 117

5.1.1 Top National Chains (2017) 117

5.1.2 Dictionary for expectation and novelty words 119

5.2 Appendix: Proofs 119

5.2.1 Proof of Lemma 2.4.1 119

5.2.2 Proof of Proposition 2.4.2 120

Bibliography **127**

List of Figures

1.1 Sentiment Analysis Methods Evaluation 10

1.2 Illustration of Attribute-Level Sentiment Analysis 14

1.3 General Architecture of a Deep Learning Network for Text Classification 21

1.4 Rating Behavior 36

1.5 Structural Model Estimates: Attribute Level Thresholds of Latent Utility by Segment 41

1.6 Odds(π) of attribute missing in reviews as a function of sentiment level . 48

1.7 Change in Average Attribute Rating 50

2.1 Followers' decisions for given prior beliefs 65

2.2 Histograms of Star Ratings: Independent versus Established National Chains 79

3.1 NSC Judgement Criteria 95

4.1 Mturk experiment: Importance of attribute sentiments 107

4.2 LDA and SLDA topics 108

4.3 Convolutional Neural Network 111

4.4 Comparison of RNN and LSTM cells (Goodfellow et al. 2016) 113

List of Tables

1.1 Examples of "hard" sentences for attribute sentiment scoring 4

1.2 Most Relevant Marketing Literature on Text Analytics 8

1.3 Algorithm for Attribute Sentiment Analysis 13

1.4 Description of Datasets 30

1.5 Class Balance: Attribute and Sentiment Classes (N: 2400) 32

1.6 Distribution of Sentence Types (N: 706) 32

1.7 Summary Statistics of Full Dataset vs. Sample 34

1.8 Sample Summary Statistics by Restaurant Type 34

1.9 Comparison of Text Mining Methods (I) 37

1.10 Comparison of Text Mining Methods (II) 38

1.11 Performance on Hard Sentence Types 38

1.12 Polarity Reversal Confusion Matrix (Sentiment Analysis) 39

1.13 Simple Confusion Matrix (Attribute Analysis) 39

1.14 Structural Model Estimates Link between Restaurant characteristics and attribute latent utility 40

1.15 Structural Model Estimates Attribute weights (normalized to sum to 1) on Ratings 42

1.16 Segment Characteristics 43

1.17 Impact of Informativeness on Attribute Presence (Segment-wise) 45

1.18 Impact of Importance on Attribute Presence (Segment-wise) 46
1.19 Model Fit: Root Mean Squared Error (RMSE) across Imputations 49
1.20 Impact of Imputation on Attribute Ratings 50

2.1 Summary Statistics of Independent and Chain Restaurants 78
2.2 Impact of Chain Dummy and Brand on Star Ratings 82
2.3 Positive, Negative and Neutral reviews by Restaurant Type 83
2.4 Presence of Expectation, Novelty and Discrepancy Words 85
2.5 Elite and Non Elite Reviewers 86
2.6 Impact of Chain Dummy and Brand on Star Ratings (For Elites) 87

3.1 Persuasive versus Non-Persuasive Videos (Content) 96
3.2 Interactivity and Textual/Audio Features 96
3.3 Most common words for each content class 98
3.4 Textual, Audio and Interactivity Features 100
3.5 Performance Comparison 101
3.6 Variable Importance 102
3.7 Variable Importance by Buyer Type (I) 103
3.8 Variable Importance by Buyer Type (II) 103

4.1 Match and Attention Comparison: Treatment (Attribute Scores) 106
4.2 Precision and Recall (Sentiment Classification) 115
4.3 Precision and Recall (Attribute Analysis) 115

5.1 Revenue, Satisfaction and Review Valence 118
5.2 Custom Dictionaries 119

Chapter 1

Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Missing Attributes

1.1 Introduction

Many firms conduct routine tracking surveys on product/service performance on selected attributes chosen by managers that they believe drive overall customer satisfaction (Mittal et al. 1999, Mittal et al. 2001). The summary scores from these surveys are used as dashboard metrics of overall satisfaction and attribute performance by managers. In many industries offering "experience goods", such as restaurants, hotels and (even) nursing homes, crowd-sourced online review platforms have emerged as an alternative and less expensive source of scalable, real-time feedback for businesses to *listen in* on their markets for both performance tracking as well as competitive benchmarking (e.g., Xu 2019, Li et al. 2019). Even when not used as a replacement for tracking surveys of performance, such quantitative summary metrics are valuable for managers because consumers use review platforms when making choices (e.g., Zhu and Zhang 2010, Luca and Vats 2013).

This paper develops a scalable text analysis method by which online review platforms

that only collect open-ended text reviews can produce attribute level summary ratings[1] similar to those who use quantitative attribute level surveys. This involves solving two novel and challenging sub-problems. First, it requires developing a text mining framework that can convert the rich texture of attribute level sentiment expressed in the text to a fine-grained quantitative rating scale, that not only captures the *valence* of the sentiment, but also the *degree* of positivity or negativity in sentiment. The second problem is that since reviewers self-select which attributes to write about in open-ended text, many attributes will be missing in unprompted reviews. The challenge is to correctly interpret "silence," when a reviewer does not mention an attribute in the review text and impute the correct sentiment to obtain the aggregate attribute level rating. Our results show that the magnitude of corrections can be large enough to be managerially significant.[2] Further, behavioral research has long recognized the importance of the right imputation for missing values because people do not ignore missing attributes and often make complex and imperfect inferences from missing data in evaluations. For example, Slovic and MacPhillamy (1974) and Peloza et al. (2015) discuss some common types of wrong inferences—higher weights on common attributes (i.e. attributes for which information is available for all options) or simply proxy missing attribute score with some unrelated attribute score (extra-attribute mis-estimation). Gurney and Loewenstein (2019) provides an excellent review of this topic. While the nature of these inferences may vary, the general takeaway is that missingness usually worsens choice and decision making. This justifies our interest in obtaining corrected attribute ratings.[3] We next describe the key challenges involved in

[1]Some review platforms such as Zagat, OpenTable and TripAdvisor ask for numerical attribute ratings from reviewers before open-ended text. This may obviate the need to convert text to numerical attribute sentiment scores; but a key disadvantage is that attribute level questions vastly reduce response rates and quality because of the additional time and cognitive costs on the reviewers (Krosnick 1991, Huang et al. 2015). Therefore many large review platforms such as Yelp, Google and Facebook only obtain an overall rating and free-flowing, open-ended text feedback. Our approach can provide attribute level ratings on such platforms.

[2]Luca (2016) finds that a 1 point change in restaurant ratings leads to a 5-9% change in revenues.

[3]To assess its value in our specific context, we show using an mTurk experiment (see Online Appendix Table A1), that consumer choices are more consistent with their true preferences when attribute level ratings are available. Managers also clearly would prefer their ratings to be valid—to the extent imputations help

tackling these two problems and how we address them.

1.1.1 Challenges in Attribute Level Sentiment Scoring from Text

Attribute level sentiment scoring from text involves connecting a specific product attribute (e.g., food, service) to an associated satisfaction rating. With fine-grained sentiment scoring, we need to convert text to more than just valence (positive, negative, neutral), but also represent the degree of positivity and negativity (in say a 1-5 point scale). While there has been some work on sentiment scoring of attribute valence (e.g., Archak et al. 2011), there has been little work on fine-grained attribute scoring—the focus of our paper. We now describe the challenges involved relative to extant work in the literature. We note that the computer science literature in fine-grained sentiment scoring is still evolving and it remains an open problem in natural language processing (Schouten and Frasincar 2015).

Over the last decade, marketing scholars have extensively used text analysis to identify topics, customer needs and mentions of product attributes. These papers typically have used "bag-of- words" approaches such as LDA and lexicons—where the identification of attributes and sentiments is based on the frequency of sentiment words. LDA applications include Tirunillai and Tellis (2014), Hollenbeck (2018), Puranam et al. (2017) and Büschken and Allenby (2016). Archak et al. (2011) use a lexicon method to identify attributes and sentiment valence; but do not address fine-grained sentiment scoring.[4] But bag-of-words based approaches are limited in their ability to adequately score attribute sentiments. Consider the following examples where *sentiment degree* is modified, as in (i) *"horrible," "not horrible," "not that horrible"* and (ii) *"delight, "just missed being a delight"*. When words are just counted as in bag-of-words, making the connections be-

obtain valid estimates, they would clearly prefer it. We show that our imputations indeed work better than other common imputations on a holdout sample, and that the corrections are large enough to be managerially significant.

[4]Timoshenko and Hauser (2018) and Liu et al. (2019) use deep learning models that are not based on word frequencies, but their focus is on attribute and valence identification respectively and hence do not need to account for language structure issues that need to be addressed in fine-grained attribute sentiment scoring.

tween the key sentiment words "horrible" and "delight" with their degree modifiers will be difficult, without considering how they are grouped adjacently to form phrases—i.e., spatial structure.

More generally, in NLP, certain types of sentences are considered "hard" for sentiment scoring (Socher et al. 2013). Table 1.1 provides a typology of such "hard" sentences with examples. Like the examples above which modified sentiment degree, *negations* often require accounting for adjacent words, i.e., spatial structure to correctly interpret both valence and sentiment degree. Further, other types of sentences such as *long and scattered sentences* and *contrastive conjunctions* require accounting for both the spatial the sequential structure of language, as the sentiment is distributed and modified across *non-adjacent words* in a sentence. When there are long sentences with sentiments scattered across attributes, being able to make the right association of the sentiment with the attribute becomes a challenge; further sentiments get modified along different parts of a long sentence, and therefore one has to consider these sequences together in inferring sentiment. Contrastive conjunctions–words/phrases like "but," "despite," and "inspite of" can reverse the sentiment of a sentence—on either side of the conjunction. Implied sentiments are challenging because the meaning/sentiment associated with a word lies within a richer context of its usage.

Table 1.1: Examples of "hard" sentences for attribute sentiment scoring

Type	**Example**
Negations and Sentiment Degree	Pizza is *good* > Pizza is *not that good* > Pizza is *not at all good*
Long sentences and Scattered Sentiments	OK, in fact *good*, to start with but *kept getting worse* and wait staff were *unapologetic* but manager *saved the night*.
Contrastive Conjunctions	Despite the *creativity in the menu*, *execution was a disappointment*
Implied Sentiments	The place is a *treasure* if *only you are lucky to be there on the right day*

These examples motivate the need to go beyond frequency-based "bag-of-words" approaches and model the structure of language (in terms of phrases and sequences). In

our deep learning model, a convolutional layer captures the spatial structure (grouping of adjacent words), and a Long Short Term Memory (LSTM) layer captures the sequential structure (sequence of adjacent and non-adjacent phrases). This allows us to improve our sentiment classification not only in the aggregate on "easy" sentences, but also on the "hard" sentences.

Accounting for Attribute Silence in Attribute Sentiment Scoring

As described earlier, the current literature on topic identification focuses on the frequency of mentions across reviews (e.g., Büschken and Allenby 2016) to identify the most common or novel needs/benefits, attributes desired by consumers/user. The implicit assumption is that topics or attributes that are not mentioned are not important and can be ignored.

We question the premise that *importance* is the primary reason for why an attribute is mentioned or not. There can be other reasons for why a reviewer is silent on an attribute. Some may write only if it can *influence* or be *informative* to readers. For example, if there is high variance among current raters, one's rating can be influential and informative. Or if one's own rating is different from the consensus based on current reviews, one may be motivated to write a different point of view. There could of course be asymmetry in this motivation depending on whether the deviation from consensus is positive or negative. Finally, some raters may choose not to write when the product meets expectations (and rating would have been a three), but only to praise/vent when they are very satisfied or dissatisfied.

We develop a model-based strategy that imputes *missing* sentiment based on observable restaurant characteristics and observable/unobservable reviewer characteristics. We consider and exploit three key features of the available data in this context in developing and identifying the structural model: (1) the same restaurant is visited and experienced by multiple reviewers; given that a restaurant provides similar services to all patrons, we

assume that all reviewers receive a common latent utility plus idiosyncratic shocks. (2) the same reviewer visits multiple restaurants, this allows us to identify observable reviewer heterogeneity and unobserved heterogeneity in rating styles—i.e. how they map experienced utility to attribute level ratings. (3) all reviewers provide an overall rating, so given multiple observations from a reviewer, we can infer heterogeneous weights of attributes on overall ratings.

We allow the structural model of rating behavior to account for heterogeneity in rating styles and weights on attributes driving overall ratings. Specifically, we allow for a non-linear and heterogeneous mapping from experienced utility to attribute ratings using an ordinal logit and a heterogeneous weighting of different attributes to explain the observed overall rating as a regression. The heterogeneity is modeled within a latent class framework. We estimate the model using an EM algorithm, where the missing data on attribute ratings are imputed based on the model parameters during an iteration and iterated till the parameters converge.

Since the structural model provides insights on reviewer segments and their behavior, it not only helps with imputation but also enables us to assess the above conjectured "drivers of silence" in reviews. We find that there are multiple reviewer segments with different motivations to write reviews—one segment seeks status, another seeks to vent/praise and a third is altruistic or wants to voice their opinion. Interestingly, we find that informativeness and need to vent/praise drives what attributes are mentioned; not attribute importance. We then validate the imputations from our structural model by showing superior performance relative to simpler homogeneous models and other ad-hoc imputation rules on holdout data. Finally, we demonstrate that corrections for attribute silence based on observable and unobservable heterogeneity leads to significant corrections in average attribute ratings for a business.

We note that our problem definition for attribute level ratings abstracts away from issues of (1) selection in *who* chooses to review (e.g., Li and Hitt 2008) and (2) strategic

review shading by reviewers and/or fake reviews (e.g., Mayzlin et al. 2014, Luca and Zervas 2016) when aggregating ratings. Reviewer selection/review shading issues are relevant not just for attribute level ratings, but also for overall ratings; as such any approaches to address these issues for overall ratings should also be applicable for attribute level ratings.

Summarizing, our key contributions are as follows: The paper is the first to do *fine-grained* attribute sentiment scoring using text reviews in marketing; i.e., we not only capture attribute sentiment valence, but also the *degree* of positivity or negativity in sentiment. For this, we highlight the need to move beyond word frequency based approaches (lexicon and LDA) to a deep learning approach that accounts for language structure. Specifically, we account for the spatial and sequential structure of language using a convolutional-LSTM model. Second, we find that attribute silence in reviews is driven by need to inform and need to praise/vent, but not based on the importance that the reviewer itself places on the attribute. Using a structural model of rating behavior, we develop a model-based imputation for missing attribute ratings. Overall, we note that though the paper is motivated in the empirical context of online reviews, the problems of generating fine-grained attribute sentiment scoring from text and the interpretation/correction of attribute silence has broad application across many settings.

The rest of the paper is organized as follows. §1.2 discusses the related literature. §1.3 describes the problem of attribute sentiment scoring, the challenges and how our model addresses these challenges. §1.4 describes the structural model of rating behavior, the estimation strategy, and how the model is used for imputing missing attribute scores. §1.5 describes our data. §1.6 summarizes the results. §1.7 concludes.

1.2 Related Literature

This paper is related to multiple strands of literature in marketing and computer science. We organize our discussion in two parts.

1.2.1 Text Analytics on UGC and Online Reviews

Table 1.2 positions our paper with respect to the most relevant literature on online reviews and user generated content in marketing. Some of the early research on user-generated (UGC) content in marketing (e.g., Chevalier and Mayzlin 2006, Dhar and Chang 2009, Duan et al. 2008, Ghose and Ipeirotis 2007, Onishi and Manchanda 2012) uses quantitative metrics like review ratings, volume and word count to infer the impact of UGC on business outcomes like sales and stock prices. Though these papers established the importance of studying UGC and its specific role in experience goods markets, they did not investigate content in review text.

Another research stream focused on using UGC content in blogs and review forums to extract insights around customer needs and brand positioning (e.g., Lee and Bradlow 2011, Netzer et al. 2012, Tirunillai and Tellis 2014, Büschken and Allenby 2016). Archak et al. (2011) use UGC to measure sentiment valence (not fine-grained sentiment) on specific product attributes using a lexicon approach and its impact on demand.

Table 1.2: Most Relevant Marketing Literature on Text Analytics

Paper	Analysis Unit	Sentiment Analysis (Y/N)	Sentiment Granularity	Method	Performance Metric	Attribute Silence (Y/N)
Godes and Mayzlin (2004) & Chevalier and Mayzlin (2006)	Doc	NA	NA	No Text Mining	NA	N
Lee and Bradlow (2011)	Doc	N	NA	Bag of Words	Overall	N
Archak et al. (2011)	Doc	Y/N	Binary	Semi-supervised	Overall	N
Netzer et al. (2012)	Doc	N	NA	Lexical Networks	Overall	N
Tirunillai and Tellis (2014)	Doc	Y/N	Binary	LDA	Overall	N
Timoshenko and Hauser (2018)	Sent	N	NA	CNN	Overall	N
Büschken and Allenby (2016)	Sent	N	NA	Sentence LDA	Overall	N
Liu et al. (2019)	Doc	Y	Binary	CNN, RNN, LSTM	Overall	N
This paper	Sent	Y	5-level	Convolutional-LSTM	Overall & Hard Sentences	Y

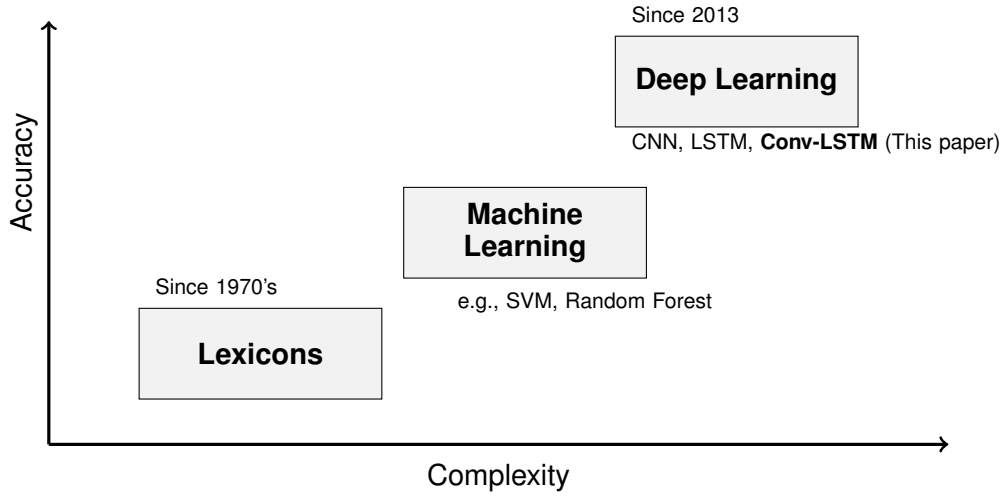
Doc: Document Sent: Sentence

Fine-grained sentiment analysis for individual attributes is one of the more challenging variants of the sentiment analysis problem (Feldman 2013, Wang et al. 2010, Nanli et al.

2012, Gonçalves et al. 2013, Keramatfar and Amirkhani 2019, Balaji et al. 2017, Qazi et al. 2017). Figure 1.1 shows the evolution of sentiment analysis literature, highlighting the trade-offs of the different approaches. Lexicon or dictionary based methods (Wang et al. 2010, Taboada et al. 2011) are highly interpretable, but rely on carefully hand-crafted features. They are therefore not scalable. They under-perform in detecting sentiments in "hard" sentences. Early supervised text classification methods like SVM (Joachims 2002) do not need hand-crafting and are scalable but they need large amounts of labeled training data (tagged by humans) to reach desired levels of accuracy. Hence deep learning models (Kim 2014, Socher et al. 2013, Zhou et al. 2015) combined with meaning-infused word vectors (Pennington et al. 2014, Mikolov et al. 2013) have revolutionized the field of text mining — they do extremely well on text classification tasks, yet require only much smaller volume of training data to attain high levels of accuracy. Thus they overcome the shortcomings of both traditional supervised as well as unsupervised algorithms. A limitation is that they lack interpretability and so it is hard to understand what is driving the performance of deep learning models. Recently, marketing scholars have used deep learning models for text analysis to answer important questions such as need identification (Timoshenko and Hauser 2018) and the impact of reading reviews on particular attributes on purchasing decisions (Liu et al. 2019), but their focus is not on fine-grained sentiment and hence language structure is less important.

We advance the marketing literature on sentiment analysis in two ways: (i) considering fine-grained attribute sentiment scoring and (ii) moving from "bag-of-words" methods like LDA and lexicons to deep learning models that account for structural aspects of language. Hybrid models that combine features of different deep learning architectures can improve performance on hard tasks (Wang et al. 2016); in that spirit, we motivate and construct a hybrid convolutional-LSTM model. Further, to understand the key drivers of model performance, we test our model on various types of hard sentences. In our corpus, nearly half of the sentences are 'hard," justifying the need to account for language structure. By

Figure 1.1: Sentiment Analysis Methods Evaluation



reporting performance metrics not just overall, but on types of "hard" sentences, we offer new benchmarks for performance evaluation in future research.

1.2.2 Missing Attributes (Attribute Silence) in Reviews

Our study of attribute silence, i.e. missing attributes in text reviews is primarily related to the statistics literature on missing data and imputations. Rubin (1976) laid the seminal framework for analysis of missing data, in which every data point has some likelihood of being missing. Rubin classifies missing data problems into three groups: "Missing Completely at Random" (MCAR), "Missing at Random" (MAR), and "Missing Not at Random" (MNAR). MCAR occurs if the probability of missing is the same for all cases, i.e., causes of the missing data are unrelated to the data. This assumption is likely violated in most settings. Most modern imputation models for missing data are based on the MAR assumption; i.e., the probability of being missing is the same within groups defined by the observed data. For this strategy to be successful, rich behavioral models (including those with unobserved heterogeneity) are modeled on the behaviors of interest, such that the MAR assumption becomes reasonable. While many MAR models are based on ob-

served heterogeneity, in our setting given the potential unobserved heterogeneity in rating styles across reviewers, the MAR strategy will be unsuccessful without unobserved heterogeneity. Fortunately, in our setting given multiple observations across restaurants and reviewers, unobserved reviewer heterogeneity can be estimated and the MAR approach can be applied. If not, then we have an MNAR setting potentially due to unobserved heterogeneity. The most common approach is to then introduce new identifying restrictions by explicitly justifying a model of missingness for the context at hand, and estimate the joint model of missingness with the behavioral model (Little and Rubin 2019, Mohan and Pearl 2018). Overall, missing data models are often estimated using Multiple Imputation, or by likelihood methods. Likelihood based approaches either use Bayesian methods or the EM algorithm for estimation. Recently, Athey et al. (2018) proposed matrix completion methods for imputation in big data settings.

In this paper, we develop a structural model of heterogeneous reviewer rating behavior that allows for both observable and unobservable heterogeneity taking into account the data generation process. We allow for a rich nonlinear mapping from experienced utility to five-level rating and weighted mapping of attribute ratings to overall rating behavior. We use an EM algorithm to estimate the model, with model-based imputation to fill in for missing attribute ratings during the EM iterations. Ex-post, we use the parameters of the structural model to assess various conjectures of attribute missingness. We also use the estimates of the structural model to impute for missing attribute ratings to construct aggregate corrected metrics of restaurant ratings, conditional on the observed characteristics of restaurants and the observed and unobserved characteristics of reviewers.

1.3 Converting Text into Numeric Attribute Sentiment Scores

We first describe the attribute level sentiment analysis problem of converting unstructured text data in reviews into attribute level sentiment scores. We then describe two methods of attribute scoring models with text data: (1) the lexicon model and (2) the deep learning model. Along the way, we also describe various implementation issues and choices that needs to be made.For completeness, we also estimate some bag-of-words based supervised machine learning models e.g., Support-Vector-Machine (SVM), Naive Bayes and Logistic Regression as baseline models as they have been used for text classification in the past. We also estimated a supervised topic model S-LDA, but do not report the results as it does not separate attribute and sentiment classes well; a primary requirement for this task. Aggarwal and Zhai (2012), Sebastiani (2002) provide a good review of these methods.

The problem of attribute level sentiment analysis is to take a document d as input (in our empirical example, a Yelp review) and identify the various attributes $k \in K$ that are described in d, where K is the full set of attributes. Having identified the attributes k, the problem requires associating a sentiment score s with every attribute. In solving the attribute level sentiment problem, we make two simplifying assumptions. First, we assume that each sentence is associated with one attribute. Occasionally, sentences may be associated with more than one attribute; in that case, we consider the dominant attribute associated with the sentence. Like Büschken and Allenby (2016), we find that in our empirical setting, multiple attribute sentences account for less than 2% of sentences in our review data, and thus have very little impact on our results. Second, we assume that the attribute-level sentiment score of a review is the mean of the sentiment scores of all sentences that mention that attribute. We outline the steps involved in obtaining attribute level sentiment ratings from text reviews in Table 1.3.

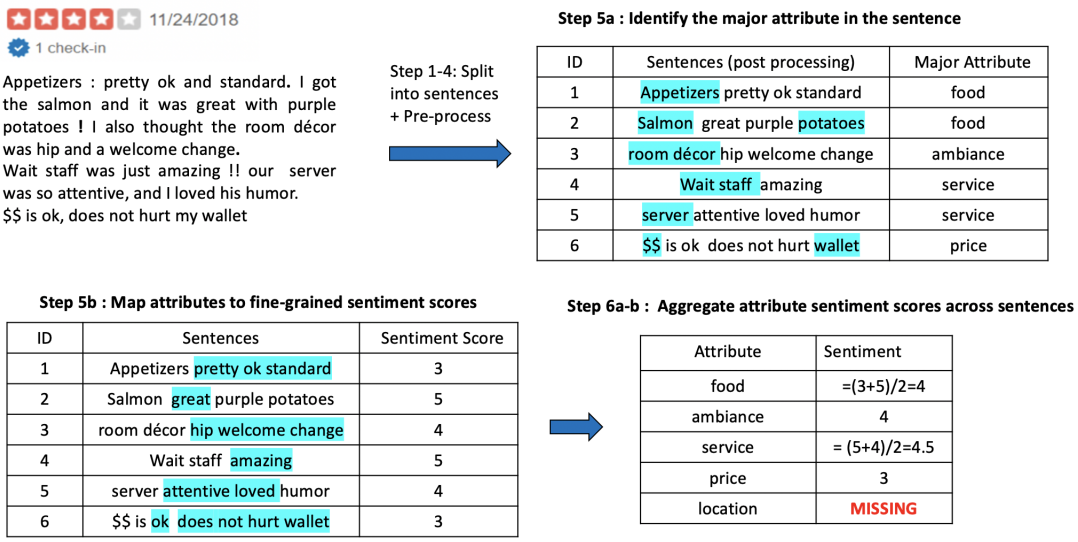
Table 1.3: Algorithm for Attribute Sentiment Analysis

Algorithm : Derive Attribute scores from Review Text
Input : Review text *s: no of sentences, w_s: words in sentence s*
Step 0 : Choose relevant sentiment and attribute scale **Step 1:** Split review doc r_d into sentence vectors of s sentences using standard tokenizers **Step 2:** For all s sentences , repeat steps 3 through 7 **Step 3:** Pre-process the sentence to convert characters to lower-case, remove stop-words and punctuations **Step 4:** Pass one sentence at a time into an Attribute Sentiment Classifier *AS* **Step 5a :** *AS* classifies sentence into an aspect class based on its algorithm **Step 5b:** *AS* classifies sentence into a sentiment class based on its algorithm **Step 6a:** Attribute Score $\longrightarrow$ mean(attribute sentiment across all sentences) **Step 6b:** If an attribute is not mentioned in any sentence s, assign it a missing sentiment score

Most of the steps in table 1.3 are clear, except for the choice of attribute/sentiment classes (Step 0) and the attribute sentiment classifier used (Step 5). We begin by describing how we choose the relevant attributes and the sentiment scale in Step 0. We use a 1-5 scale for sentiment granularity (1: extremely negative, 3: neutral and 5: extremely positive) as this is comparable to the 5 point rating scale in many review platforms. Also, human taggers fail in practice to differentiate well between classes when the sentiment granularity is higher than 5 levels (Socher et al. 2013).

To obtain comparable fine-grained sentiment scores on a managerially relevant set of attributes across restaurants, we first need to choose a set of attributes on which restaurants should be scored on. This is similar to an exploratory phase before conducting a quantitative survey. For this, we conducted (i) a review of the literature; (ii) an analysis of the most frequent attribute words in the corpus; and (iii) topic modeling using LDA. The literature on restaurant evaluation and industry customer satisfaction surveys identified food quality, employee behavior and wait time (service), basic hygiene, look and feel (ambiance) and value for money as the most common attributes (Ganu et al. 2009). We then did frequent word categorization of our review corpus by associating the most high frequency nouns, noun phrases and select verbs to restaurant-relevant attributes. Be-

Figure 1.2: Illustration of Attribute-Level Sentiment Analysis



yond the four attributes identified from past literature and industry surveys, we found a fifth attribute "location" that has words pertaining to parking, convenience and safety of the restaurant location. Finally, we conducted topic modeling of our review corpus using LDA. As is common with LDA, these topics combined both restaurant attributes and consumer sentiments, and given the very high frequency of food related comments, the topics were disproportionately around food.Büschken and Allenby (2016) note that by initializing the LDA model with seed-words for a wider range of attributes, one could obtain more balanced topics. Since we only needed to identify relevant topics and not gain greater balance, using seed-words did not help with identifying additional attributes that were relevant for a large enough set of restaurants to be used on a platform Overall, we concluded that the five attributes—food, service, ambiance, value and location captured the most relevant attributes for a restaurant rating platform.

Figure 1.2 illustrates and clarifies the major steps in attribute sentiment scoring using an example review. These steps above are the same irrespective of the Attribute Sentiment Classifier (*AS*) used in Step 5 of Table 1.3. We next describe the two types of attribute sentiment classifiers we consider.

1.3.1 Attribute Sentiment Classifier: The Lexicon Method

We begin with the lexicon-based method because it is highly interpretable, transparent and very widely used and thus serves as useful benchmark relative to more complicated models. The method consists of lexicon construction followed by attribute sentiment classification of text based on dictionary look-ups; i.e. sentences are classified into an attribute and sentiment class by locating word matches in attribute and sentiment class-specific dictionaries. We explain the method below and discuss its limitations.

1. Lexicon building. Lexicon construction involves creating a dictionary of attribute words with corresponding attribute labels (e.g., *waiter*–"service") and sentiment words with sentiment class labels (e.g., *excellent*–"extremely positive"). We first identify the high-frequency attribute and sentiment words in our corpus to create our *vocabulary*. We construct attribute and sentiment class-specific dictionaries, by asking human taggers on Amazon Mechanical Turk to classify all attribute words into one of the five attributes we identified in Step 1—food, service, value, ambiance and location and all sentiment words into one of the five sentiment classes–given we decided to use a 5 point rating scale. Every word is labeled by 3 distinct human taggers and we retain only those words for which at least 2 out of 3 taggers agree on the labeling.[5]

2. Attribute Level Sentiment Scoring. Each review is split into sentences. Using the lexicon, each attribute word in the sentence is classified into one of the pre-specified attributes (or none) and each sentiment word is classified into a 1-5 sentiment rating scale using a "look-up" or search of the pre-created lexicons. Following this, the steps are similar to those listed in Table 1.3

[5]While it is possible to use a previously constructed generic lexicon to label attributes and to assign sentiment scores, a domain and task specific lexicon improves classification/labeling accuracy. Moreover, we could not find any existing lexicon that is well-suited for fine-grained sentiment analysis of restaurant reviews. For e.g., AFINN lexicon (Nielsen 2011) and Stanford Sentiment Treebank (Socher et al. 2013) have words and phrases with 5-levels of sentiment classification, however, they are built on Twitter and rotten.tomatoes.com movie review dataset respectively and have limited overlap of words and attributes with our restaurant domain.

Despite its simplicity, interpretability and transparency, the method has several limitations. First, lexicon construction is costly in both time and effort, and scales linearly with number of words. Second and more importantly, the method treats language as simply a bag-of-words or "fixed phrases" and does not account for various aspects of language structure. In practice, lexicon methods therefore work fairly well for sentiment identification in simple sentences, but perform poorly on "hard" sentences (Liu et al. 2010).

Why the Lexicon Method Fares Poorly with "Hard" sentences. We elaborate further on why lexicon methods fail to classify *hard* sentences that we had mentioned in the introduction in Table 1.1 (Socher et al. 2013). This is problematic because "hard" sentences are close to 50% of sentences in our review corpus. We now explain each of these types.

1. Negations and Sentiment Degree. Sentences which have different degrees of negative sentiment can be hard to classify without accounting for variable size n-grams. Lexicon methods typically look at one word at a time and will not be able to obtain sentiment valence or degree; Even if ad hoc approaches may be used to address standard negations with bi-grams or tri-grams by hard-coding negation phrases, examples like "*Pizza is not that good*," "*Pizza is not at all great*," illustrate that such ad hoc approaches are unlikely to be effective overall in capturing degree of sentiment. This motivates the use of the convolutional layer, which handles the spatial structure.

2. Long Sentences and Scattered sentiments. In long sentences consisting of more than 20 words, the degree of sentiment (and even polarity) can change multiple times. As an example," *OK, in fact good, to start with but kept getting worse and wait staff were unapologetic but manager saved the night*." In this sentence, the sentiment flows from being good to bad to extremely bad and then back to positive. Yelp reviews tend to have a significant percentage of long sentences. Without sequence history, the classifier cannot capture sentiment shifts and will classify most of these sentences as *neutral* due to the mix of positive and negative sentiment words. More importantly, immediate sentiment modifiers may be changed by sentiment words that are farther away, so having a "long

term memory" of what was said before and whether recent sentiment (short-term memory) should take precedence needs to be considered. The LSTM layer helps with both the sequencing and the immediate and distant sentiment modifiers, while the convolutional layer still helps group words into phrases within the long sentence before being fed into the LSTM layer.

3. Contrastive conjunctions. Sentences which have an *X but Y* structure often get misclassified by sentiment classifiers as the model needs to take into account both the clauses before and after the conjunction and weigh their relative importance to decide the final sentiment. An example sentence includes "*Despite the creativity in the menu, execution was a disappointment*." The first half here is extremely positive due to the word *creativity*, but the second half moderates it significantly. A good classifier should be able to learn from both parts of the sentence to arrive at the correct classification. While the convolutional layer identifies phrases before and after the conjunction, the LSTM layer helps with interpreting the change of meaning after the conjunction.

4. Implied sentiments (sarcasm and subtle negations). These sentences do not have explicit positive or negative sentiment words but the context implies the underlying sentiment. This makes the task of sentiment identification extremely hard for all classes of models and especially for models relying on a specific set of positive or negative words. An example sentence includes "*The place is a treasure if only you are lucky to be there on the right day*." This is an example of sarcasm, the reviewer uses a positive word like "treasure" but hints at the extreme variance in the type of experience one can have. There could also be subtle negations, for example, "*The girl managing the bar had to be the waitress for everyone*." Here the reviewer is complaining about lack of service arising out of shortage of staff without using any explicit negative word. Given the meaning/sentiment associated with the work lies in the richer context of its usage, we will empirically assess how much the spatial and sequential structure helps with accurate classification.

1.3.2 Attribute Sentiment Classifier: A Deep Learning Hybrid Convolutional-LSTM Model

Lexicon methods use a constructive algorithm based on pre-coded attributes and sentiment words in a lexicon to score attribute level sentiment. In contrast, deep learning models are a type of supervised learning model, where the model is trained using a training dataset by minimizing a loss function (e.g., the distance between the model's predictions and the true labels). The trained model is then used to score attribute level sentiment on the full dataset. Like deep learning, regression and support vector machines (SVM) are also variations of supervised learning.

What distinguishes deep learning from regression and support vector machines is that deep learning seeks to model high-level abstractions in data by using multiple processing layers (the multiple layers give the name "deep"), composed of linear and non-linear transformations (Goodfellow et al. 2016). Deep learning algorithms are useful in scenarios where feature (variable) engineering is complex and it is hard to select the most relevant features for a classification or regression task. For instance, in our task of fine-grained sentiment analysis, it is not clear which features (combination of variable length *n*-grams) is most informative in order to classify a sentence into "good food" or "great service". The two key ingredients behind the success of deep learning models for NLP are meaningful word representations as input and the ability to extract contiguous variable size n-grams (spatial structure) with ease while retaining sequential structure in terms of word order and associated meaning.

In this section, we outline the architecture of the model and its intuition and discuss critical modeling/implementation choices.[6] Figure 1.3 shows the general architecture of a neural network used for text classification. Following pre-processing of text, the first layer is the embedding layer, where words are converted to numerical vectors by making use

[6]The technical description is provided in a self-contained online appendix for the interested reader.

of word embeddings. These embedded numerical vectors are then fed to the succeeding feature generating layers, which are the core of the deep learning model. In contrast to older supervised learning methods like SVM which work with the raw data directly as inputs, these feature generating layers, i.e., the convolutional layer and long short term memory network (LSTM) layer in our model, extract higher level features important for classification. The extracted feature vectors are then passed into a logit classifier (softmax) that classifies the sentence to the class with highest probability of association.

Embedding Layer and Word Representation. Neural network layers work by performing a series of arithmetic operations on inputs and weights of the edges that connect neurons. Hence, words need to be converted into a numerical vector before being fed into a neural network.The simplest method to form numeric vectors from words is a one-hot representation which means that if there are V words in the vocabulary; each word is represented as a $V \times 1$ dimensional vector where exactly one of the bits is 1 and rest are zero. Such a representation is not scalable for large vocabularies and also stores no semantic information about words. Another option is to only take into account word frequency and simply convert words into numbers based on some normalized frequency score like *tf-idf*. These vectors are called *embedding* and most well-known embedding algorithms (e.g., word2vec, GloVe) are based on the distributional hypothesis— words with similar meanings tend to co-occur more frequently (Harris 1954) and hence have vectors that are close in the embedding space. The efficiency of the neural network improves manifold if these initial inputs carry meaningful information about the relationships between words. Hence the choice of embedding is an important one — we experiment with both embeddings trained from scratch on our Yelp review corpus as well as a range of pre-trained word embeddings like Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014) that are available for all words in our *vocabulary*.These embeddings have been trained on different corpus like Wikipedia dumps, Gigaword news dataset and web data from Common Crawl and have more than 5 billion unique tokens. There are pros and cons for both approaches

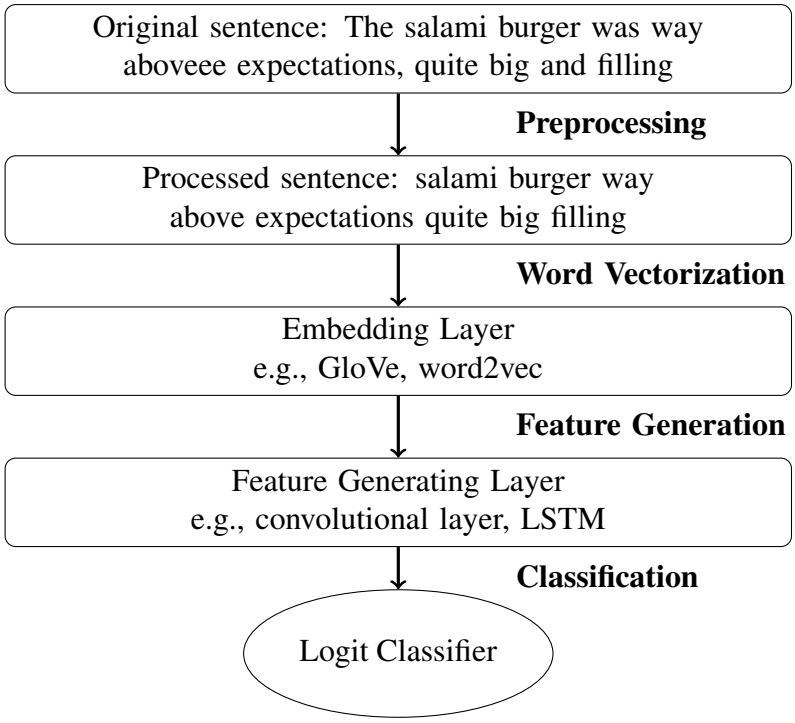
— pre-trained embeddings is a form of transfer learning that eliminates embedding generation time, but self-trained embeddings may result in higher classification accuracy due to a more context-relevant vocabulary.

Feature Generating Layers (Convolutional-LSTM). The macro architecture of the neural network comprises of layers to be included (e.g., *feed-forward* or *convolutional*) and type of interconnections between them. As discussed above, the most challenging aspect of our task is dealing with different types of hard negations resulting from variable-size n-grams (e.g., *not good, not that great*) and shifting polarities (*started off well but ended in a sorry surprise*). In many challenging text and image classification problems (Wang et al. 2016), hybrid models that combine the strengths and mitigate the shortcomings of each individual model have been found to improve performance. In that spirit, we build a network consisting of a single convolutional layer with variable-size filters followed by a Long Short Term Memory (LSTM) layer.

Convolutional layers with different filter sizes specialize in extracting variable-length n-grams (phrases) associated with relevant attributes and sentiments and have recently been used successfully in various text analysis applications (Kim 2014, Timoshenko and Hauser 2018). To improve granular sentiment detection where sequence information is critical, we follow the convolutional layer with an LSTM layer that processes the features (phrases) identified from the convolutional layer. LSTM is a variant of the recurrent neural networks (RNN) that specializes in handling longer contextual information (Hochreiter and Schmidhuber 1997). An LSTM employs a cell state (long-term memory) and a combination of gates that are like "regulators" of information to constantly evaluate what parts of the history (in this case n-grams from earlier part of the sentence) need to be forgotten and what needs to be retained to improve the accuracy of the attribute and sentiment classification task.For more details on this architecture, see online appendix. As we motivated in our discussion of "hard" sentences, by taking advantage of the properties of the convolutional layer and LSTM, we expect the hybrid to improve classification accuracy while

keeping training time low.

Figure 1.3: General Architecture of a Deep Learning Network for Text Classification



Classifier. The loss function choice depends on the nature of the classification task. Since our tasks involve the classification of text into 5 attribute classes and 5 sentiment classes, it is a multi-class classification problem. We use the standard loss function for multi-class classification called Categorical Cross Entropy. Say s_i represents the convolutional-LSTM model classification for sentence i and t_i represents the ground truth classification, then the cross entropy loss function can be defined in the following manner :

$$Categorical\ Cross\ Entropy\ Loss\ (CCE) = -\sum_{i}^{C} t_i log(s_i)$$

1.3.3 Deep Learning Implementation: Important Choices

Word Embeddings. We tested pre-trained embeddings based on word2vec and GloVe with different numbers of embedding dimensions (e.g., 100, 300) for attributes and sentiment

classification. Further, we evaluated whether self-trained embeddings from the specific text corpus can produce superior classification relative to the pre-trained embeddings.

Micro Architecture. The micro architectural decisions in a neural network involve the number of neurons in each of the layers, the size and number of filters for the convolutional layer and dimensions of the max pooling function (that concatenates variable-size feature vectors generated from variable-size convolutional filters). Many of these decisions are empirically driven but some factors that inform these choices are: sentiment classification would rely on presence of long-range n-grams, so we would typically chose a mix of filter sizes for this task ranging from 1-6 grams. In contrast, the attribute classification task often needs only unigrams and bi-grams (*chicken, cola drink, wait time*) and hence simple unigram and bigram filters would be sufficient. Also, since the sequence of n-grams matters for sentiment classification, ideally we should not use a max pooling layer after the convolutional layer as the aggregation loses sequential information before being passed to the LSTM layer. However, a pooling layer is needed to merge variable-size feature maps generated from the convolutional filters. We balance this tradeoff by max-pooling on the smallest possible pooling dimension so that we can preserve as much of the sequence information as feasible in sending input into the LSTM layer.

Model Training. As is standard for deep learning models, the model parameters are optimized jointly by training the model iteratively on smaller sub-samples of the training data (mini-batches) and then using the estimation error to improve the model (i.e. change the weights and biases in small increments) through a feedback loop. We experimented with mini-batch sizes of 5, 10, 25, 30, 50 and different optimizers. We chose the RMSProp (Bengio and CA 2015) optimizer because it uses an adaptive learning rate.

1.3.4 Performance Measures for Model Comparison

The primary metric on which we compare our models is accuracy or hit rate. This metric is formally defined as:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{1.1}$$

where tp, tn, fp, fn stand for true positives, true negatives, false positives and false negatives respectively. Accuracy is the most common metric that is used for evaluating granular text classification problems and is a fairly good metric unless there is a class imbalance issue (i.e. some classes are not well-represented in the training or test dataset). While we try to maintain class balance in our data sets, equal representation of all classes is difficult as some classes like food, service appear much more often in Yelp reviews than other classes. Likewise, moderately positive sentiments are more common than extremely positive or negative sentiments.

Among the models that do equally well on accuracy, we further evaluate them based on two types of accuracy metrics that capture not just error-rate but also the type of errors that occur.

Simple Confusion Matrix for Attribute Classification Accuracy: This confusion matrix helps to evaluate class-wise accuracy—doing so allows us to assess whether overall higher accuracy comes only from superior performance in high high-occurrence classes like food or a class like location that has few attribute words. We can assess whether the model is able to capture more complex classes like ambiance and service which manifest with a varied set of attribute words.

Polarity Reversal Confusion Matrix for Sentiment Accuracy: Though the CS literature typically uses accuracy as a performance metric for the fine-grained (multi-class) sentiment classification (Socher et al. 2013, Kim 2014), there can be other useful metrics of performance. For example, it may be useful to construct a polarity based coarse class:

positive, neutral, negative and assess accuracy on the coarse classes classification because confusing sentiment class 1 with 4 or 5 (a polarity reversal) is worse than confusing 1 with 2 (same polarity). With this thought, we construct and report a polarity reversal confusion matrix for the models that have the best overall accuracy.

We also evaluate model performance on specific hard sentence types (e.g., long and scattered sentiments, contrastive conjunctions and implied sentiments) that we discussed earlier in motivating why we account for the spatial and sequential structure of language . Finally we also assess qualitative factors like model building effort, scalability and interpretability for the various models.

1.4 Analysis Of Structured Ratings Accounting For Missing Attributes

In the first part of the paper, we converted review text into numerical attribute scores on a 1-5 scale and attributes were coded as "missing" when reviewer is silent on an attribute. For every review, we also have an overall rating on the restaurant. The challenge is how to impute the missing attribute ratings to obtain the correct aggregate attribute rating. We now outline our model-based imputation strategy to correct for attribute silence, before providing specific details.

We first develop a structural model of rating behavior that allows for (1) nonlinear mapping from experienced quality to attribute ratings; (2) heterogeneity in rating styles; and (3) heterogeneity in weights of attribute ratings on the overall restaurant rating. We then use an iterative two step EM algorithm to estimate the model, where the mapping from experienced quality to the nonlinear, heterogeneous attribute rating is estimated in the first step, and the heterogeneous weights that link attribute rating and overall rating is estimated in the second step. In each iteration, when attribute rating is missing in the

review, we impute the attribute rating based on the model estimates in the current iteration. We iterate till the model converges.

The structural model estimates give us insights into reviewer segments and reviewer rating behaviors. We use the estimates to assess whether our conjectures on attribute silence has support in the data. We then assess the validity of the model-based imputations on a holdout sample. Finally, we illustrate that corrections for attribute rating using the imputations can be substantial.

1.4.1 A Structural Model of Rating Behavior

Every reviewer who writes a review has an experience with the restaurant. Let A^*_{jk} be the experienced latent quality at restaurant j on the attribute k. The experienced latent utility is a function of observable restaurant characteristics associated with the attribute X_{jk} and an idiosyncratic shock that varies across visits.This experienced quality can vary over time t as a function of observable restaurant characteristics that vary over time, but for simplicity of notation, we suppress the t subscript in the exposition. Specifically, let

$$A^*_{ijk} = \alpha_k X_{jk} + \nu_{ijk}$$

where ν_{ijk} follows a Type I extreme value distribution (TIEV).

The mapping from underlying latent utility A^*_{jk} to the 5 point rating scale A_{jk} can be nonlinear and heterogeneous across reviewers in terms of both observable and unobservables. Specifically, we formulate the nonlinear mapping from latent experienced utility A^*_{ijk} to an ordinal rating A_{ijk} (1-5 scale) as an ordinal logit model, given that we assume ν_{ijk} to be TIEV:

$$A_{ijk} = s, \text{ if } C^g_{k(s-1)} < A^*_{ijk} + \beta^g_k X_i \leq C^g_{ks} \tag{1.2}$$

where $C^g_{k(s-1)}$ and $C^g_{k(s-1)}$ are the cutoffs of reviewer segment g for attribute k, score s

($C^g_{k0} = -\infty, C^g_{k5} = \infty$). The thresholds C^g_{ks} increase monotonically over s. While X_i captures the effect of observable characteristics on thresholds, the cut thresholds C^g_{ks} can capture the unobserved heterogeneity in reviewer's attribute rating style (for high and low scores) and differences with respect to attribute expectations, which determine satisfaction.

Further, we observe the overall rating of the restaurant for all reviews. It is natural to treat the overall restaurant rating as arising from a weighted sum of the ratings on attributes, allowing for both observable and unobservable reviewer heterogeneity (by same latent class as for attribute ratings). Specifically, we formulate the ratings equation as

$$R_{ij} = \gamma^g_0 + \sum_k \gamma^g_k A_{ijk} + \varepsilon_{ij} \tag{1.3}$$

1.4.2 Model Estimation and Missing Attribute Ratings

To the extent that there are no shared parameters across equations (1.2) and (1.3), the two equations can be estimated independently. However, given the unobserved heterogeneity, the model needs to be estimated using an iterative two step EM algorithm. Each equation is estimated in a separate step, then the posterior of the heterogeneity distribution is obtained using Bayes rule, and the iterations continue conditional on the posterior heterogeneity from the previous step until there is convergence in the heterogeneity classification of the reviewers.

In our setting, where reviewers are silent on several attributes, A_{ijk} is missing in many reviews for many attributes. In the absence of unobserved heterogeneity, imputation would be a straightforward prediction based on observable restaurant and reviewer characteristics. However since A_{ijk} is also a function of unobserved characteristics of reviewer i, the imputation needs to condition on the unobserved heterogeneity and iterated through the EM algorithm. Specifically, we use the prediction from the first step (ordinal logit), conditional on the unobserved heterogeneity as the imputation of attribute rating in the second

step and iterate till convergence.

The predicted probability of the attribute score being s ($s \in 1,2,3,4,5$) from the ordinal logit model of first step is

$$p^g_{ijks} \equiv Pr(A_{ijk} = s) = Pr(C^g_{k(s-1)} < A^*_{ijk} + \beta^g_k X_i \leq C^g_{ks}) \tag{1.4}$$

Based on the estimated β^g_k and C_s, we compute probabilities of each attribute rating for each k and s for each latent segment g when attribute rating is missing.

We then estimate the rating equation with imputation when attribute ratings are missing:

$$R_{ij} = \sum_g q^g_i \left[\gamma^g_0 + \sum_k \gamma^g_k \left[\underbrace{(1 - M_{ijk})}_{\text{if present}} A_{ijk} + \underbrace{M_{ijk}}_{\text{if missing}} \underbrace{\sum_{s=1}^{5} (s p^g_{ijks})}_{\text{Expected}} \right] \right] + \varepsilon_{ij} \tag{1.5}$$

where R_{ij} is reviewer i's star rating for restaurant j, M_{ijk} is whether the rating for attribute k is missing. If the attribute rating is present (i.e., $M_{ijk} = 0$), we use observed attribute rating A_{ijk}, and otherwise, we use expected attribute rating $\sum_{s=1}^{5}(sP^s_{ijk})$ as the input. We estimate intercept γ^g_0, attribute importance γ^g_k and probability of reviewer i belonging to segment g, q^g_i. Thus, the parameters to be estimated are $\Theta = \{\alpha_k, C^g_{ks}, \gamma^g_0, \gamma^g_k, q^g_i\}$

To be specific, the EM estimation procedure is the iteration between E (Expectation) and M (Maximization) steps below.

1. Initialization: Determine initial value of parameters $\Theta^{(1)}$ through MLE by assuming no unobserved heterogeneity across reviewers. Assume that each reviewer is equally likely to be in each segment (i.e., $q^g_i = \frac{1}{N_g}$, where N_g is the number of segments).

2. E step: For reviewer $i = 1,2,...,m$, given the nth parameter $\Theta^{(n)}$, compute $p^{g(n)}_{ijks}$, the predicted probability of the attribute score of each review.

3. M step: Estimate $(n+1)^{th}$ parameters $\Theta^{(n+1)}$ by iteratively maximizing the likeli-

hoods in step 1 and step 2.

(a) Step 1: Attribute Rating

$$\sum_{i}^{m} log(L(\alpha, C_s, \pi)) = \sum_{i}^{m} log(\sum_{g} \pi_g L_{ig})$$

where individual reviewer's likelihood $L_{ig} \equiv L_i(\alpha, C_s^g) = \prod_j \prod_l^4 \left[p_{ijk,l-1}^g - p_{ijkl}^g \right]^{1(s_{ij}=l)}$ and π_g is the segment size.

(b) Step 2: Overall Rating

$$\sum_{i}^{m} log(L(\alpha, C_s, \pi)) = \sum_{i}^{m} log(\sum_{g} \pi_g L_{ig})$$

where $L_{ig} \equiv L_i(\gamma_0^g, \gamma_k^g) = \prod_j \phi(\varepsilon_{ijk} | \gamma_0^g, \gamma_k^g)$

4. Iterate between E step and M step until convergence.

The E step is internally consistent because the imputation is based on observed and unobserved heterogeneity conditional on estimates of every iteration. Our imputation strategy works because we have multiple observations on attribute ratings and overall reviewer ratings—even if some of the attribute ratings are missing. We are able to identify the heterogeneity in attribute rating styles (the unobserved thresholds C_{ks}^g) as long as we have variation in attribute ratings on a subset of reviews from every reviewer, conditional on latent experience which are identical across reviewers and vary only by restaurant observables.

1.4.3 What Drives Attribute Silence?

We conjecture several possibilities for why a reviewer may be silent about some attributes in a review: Informativeness, Importance, and Praise/Vent need.

1. *Informativeness*: Reviewers write on review platforms to share their experience with others, so one of the major motivations could be to inform or add new information (Berger 2014). For instance, price and location may be written about less because they are not only search attributes, but prices are usually described categorically on review platforms. Location information may be often obtained through the address and other information. Hence such attributes may be described less overall. But even with experience attributes like food, service and ambiance, there may be variations in motivations across restaurants and time. For example, a review may be informative if there is high variance in predicted utility for a restaurant, i.e., there is high variance in past reviewer ratings. Controlling for variance, if the restaurant's average attribute rating is very different from the reviewer's corresponding rating, the reviewer may also consider it informative to write a review. Empirically, we assess the conjecture testing whether attribute presence (silence) is positively (negatively) related to (i) variance in past reviewer ratings, and (ii) difference between the restaurant's average attribute rating and the reviewer's rating. Further, we test whether there is heterogeneity for positive and negative deviations.

2. *Importance*: A reviewer may be silent about an attribute if it is unimportant for the reviewer. To assess if attribute silence may vary by its importance on overall ratings across unobserved segments, we empirically assess whether attribute presence (silence) is correlated with attribute weights (γ_k^g) derived from the structural model controlling for attribute type.

3. *Praise/Vent Need*: Some reviewers may feel the need to praise/vent, when highly satisfied or dissatisfied, but not write when the rating is average (when it is a three). For this we assess whether silence varied by attribute rating level. To assess this conjecture, we compare the probability of the attribute score being s when the attribute is missing ($Pr(A_{ijk} = s|M_{ijk} = 1)$ versus when it is present ($Pr(A_{ijk} = s|M_{ijk} = 0)$. Let us define a ratio $\pi_{gk}^s = \frac{Pr(A_{ijk}=s|M_{ijk}=0)}{Pr(A_{ijk}=s|M_{ijk}=1)}$. If the ratio is larger than 1 (i.e., the probability is larger in the case of missing), a reviewer who evaluates the attribute as score s is more likely to miss

the attribute. In other words, π^s_{gk} captures how likely the attribute's true score is s when it is missing vs. present. For segment g and attribute k, π^l_{gk} should be larger than 1 when score l represents satisfaction level that is more likely to be missing.

1.5 Empirical Application

1.5.1 Data

Yelp is a crowd-sourced review platform where reviewers can review a range of local businesses e.g., restaurants, spas & salons, dentists, mechanics and home services to name a few. The website was officially launched in a few U.S west coast cities in August of 2005 and subsequently expanded to other U.S cities and countries over the next few years. As of Q1 2017, Yelp is present in 31 countries, with 177 million reviews and over 5 million unique businesses listed (Yelp Investor Relations Q4 2018). Given our empirical application, we focus on restaurant reviews. Since 2008, Yelp has shared review, reviewer and business information for select U.S and international cities as part of its annual challenge. Unique reviewer and business identification numbers in the data helps create a two-way panel of reviews at reviewer and business level. For each review, we observe overall rating, textual evaluation and date of posting as well as information about business characteristics (e.g., cuisine, price range, address, name) and reviewer characteristics (e.g., experience with Yelp, Elite membership). Table 1.4 summarizes the various data sets we use for different types of analysis. A discussion on each dataset follows.

Table 1.4: Description of Datasets

Data	**Size**	**Criteria**	**Purpose**
Yelp Restaurant Corpus	1.2 Mn reviews	All restaurant reviews	Exploratory Analysis
Supervised Learning	2400 sentences	Balance of attribute and sentiment classes	Training/Testing Supervised Models
Stratified Sample	45,652 reviews	Business$\geq$20 reviews Mix of Business and Reviewer Types	Estimating Structural Model
Restaurant Panel	250,000 reviews	Restaurants in Stratified Sample	Deriving past review characteristics

1. Exploratory Analysis. We use the full dataset of 1.2 million restaurant reviews for the exploratory analysis to identify attribute and sentiment classes that we described in the model section. We created a *vocabulary* of 8458 words consisting of both sentiment and attribute words.We excluded stop-words, meaningless phrases and the long tail of words with occurrence frequency less than 1500 in our corpus. We then did a Parts of Speech tagging of our word list i.e. we classified our word list into adjectives, adverbs, nouns and verbs so as to separate attribute and sentiment words. Attribute words are mainly nouns whereas sentiment words are adjectives and adverbs with some important exceptions: for instance, some verbs are strong indicators of an attribute. e.g, "greeting", "seated", "served" refer to *service* and "spent" refers to *value*. Some adjectives are good indicators of both attribute and sentiment for e.g. the word "cheap" invariably refers to price attribute in a negative way whereas some descriptive adjectives strongly refer to an attribute for e.g., decorated refers to ambiance. Finally human taggers classified the attribute and sentiment words into attribute and sentiment classes. In our dictionaries, we only retain those words that have been labeled into a particular class by at least 2 out of 3 taggers.Our attribute and sentiment dictionaries are available upon request. These are more detailed relative to previous studies (Pak and Paroubek 2010, Berger et al. 2010) that focus on two (i.e. positive and negative) or three levels (i.e. positive, neutral and negative) of sentiments.

2. Training and Test Data for Supervised Learning . For supervised learning, we constructed another data set at the sentence level. Human taggers classify the sentences into its primary attribute and sentiment level. We ensured this dataset of sentences is balanced in its representation of all attribute and sentiment classes. 75% of this data was used for training and the remainder for model validation and testing. See Table 1.5 for the composition of training and test data sets.

An discussed in §1.3, lexicon methods cannot deal with hard sentence types. Table 1.6 shows the distribution of different sentence types in a randomly sampled subset of

Table 1.5: Class Balance: Attribute and Sentiment Classes (N: 2400)

Attribute			Sentiment		
Class	Training Data	Test Data	Class	Training Data	Test Data
Food	34%	37%	Negative	18%	25%
Service	21%	23%	Positive	35%	27%
Ambiance	14%	12%	Very Negative	11%	9%
Value	11%	10%	Very Positive	21%	27%
Location	4%	7%	Neutral	15%	11%

sentences from our corpus. 48% of all sentences and 66% of the negative sentences belong to one of the complex types. Long sentences account for 27% of our data. Given their empirical importance, we created a special test dataset of hard sentence types to assess model performance specifically on such sentence types.

Table 1.6: Distribution of Sentence Types (N: 706)

	Positive	Neutral	Negative	
Overall	52%	12%	36%	
Simple	64%	53%	34%	52%
Implied	6%	5%	32%	15%
Contrastive	7%	20%	11%	10%
Long	26%	24%	28%	27%

3. Restaurant and Reviewer Stratified Sample. To estimate the linkages between attribute level sentiment and overall ratings, we focus on a stratified sample of reviews. We ensure that we have multiple reviews by individuals so that we can account for unobserved heterogeneity in reviewer rating styles. We want multiple reviews on restaurants to ensure that there are multiple reviewers who obtained similar latent utilities up to a random shock. We therefore restricted our sample to only individuals that posted at least 5 reviews and restaurants that have at least 20 reviews.The restriction of 5 or more reviews also allows us to eliminate human or bot-generated fake reviews, which are mostly generated by users with one or only a few number of reviews. Luca and Zervas (2016) document that a larger number of reviews by a Yelp user is negatively correlated to the probability of his reviews getting filtered as spam by Yelp.

We then used stratified sampling by restaurant and reviewer types to ensure that various groups of restaurant types (high and low end; chain and independent) and different types of reviewers (elite and non-elite; by experience on Yelp) are represented in the data. This allows us to study how ratings and missing attributes differ by the types.

The sampling leaves us with 45,652 reviews from 2,704 businesses and 19,583 reviewers. As past restaurant reviews might impact current reviews, we incorporate restaurants' time-varying features (e.g., variance and mean of past reviews) by extracting all past reviews for the restaurants in our stratified sample. The full dataset (including all past reviews for restaurants in our sample) contains 250K reviews. We generate each review's time varying variables, including number of past reviews; mean and variance of past star rating; and mean and variance of past attribute ratings.

Table 1.7 compares the descriptive characteristics of the full data and our final sample consisting of 45,652 reviews. The mean and median number of reviews per reviewer in our sample is slightly higher than the population (due to stratification). However, the reviewers in our sample are fairly similar to the population in terms of average star rating, experience and length of reviews. Table 1.8 provides the number of businesses, reviews and the summary of star rating by a restaurant's price range, chain/independent Our sample has almost an equal mix of chain and independent restaurants but independent restaurants get more reviews with higher ratings on average. Low-end and high-end restaurants do not show much difference in terms of average star rating.

1.5.2 Descriptive Evidence on Attribute Rating Behavior

We now present descriptive evidence on potential drivers of reviewer's rating behavior to motivate the choices and assumptions we make in the structural model. We first look at the impact of observable reviewer (e.g., Elite status[7]) and restaurant characteristics (e.g.,

[7]Elite reviewers receive an *Elite* badge that is displayed on their profile. They also get invited to special events. Most other observable characteristics are highly correlated with Elite, for e.g., elites are generally

Table 1.7: Summary Statistics of Full Dataset vs. Sample

	Full			Sample		
Number of Reviews	1.2M			45,652		
Number of Reviewers	1.02M			19,583		
	Mean	Median	SD	Mean	Median	SD
Star Rating	3.7	3.8	1.09	3.6	3.76	0.92
Number of Reviews per Reviewer	24	5	82	25.15	17	23.2
Reviewer's Experience on Yelp	58	56	27.5	54.6	51.8	36.2
Review Length (number of characters)	1,109	599	732	709	498	670

Table 1.8: Sample Summary Statistics by Restaurant Type

	All	By Price Range		By Chain	
		Low-end	High-end	Chain	Non-Chain
Number of Businesses	2,707	1,611	1,096	1,063	1,644
Number of Reviews	45,652	21,066	24586	10,528	35,124
Star Rating: Mean (SD)	3.5 (1.4)	3.4 (1.4)	3.6 (1.4)	2.8 (1.5)	3.7 (1.3)

price range, chain restaurant[8]) on the distribution of attribute ratings and attribute *missingness*. Figure 1.4a shows differences between the rating behaviors of Elites and Non-Elites. X-axis represents each rating or missing indicator. Elites tend to give more moderate ratings (3 and 4 stars) whereas Non-Elites give more extreme ratings (2 and 5 stars) across attributes. More importantly, Non-Elites tend to *miss* more attributes in their reviews (especially *ambiance*). Such differences suggest Elites might have different motivation to give ratings than Non-Elites.

Figures 1.4b and 1.4c show how rating behavior differs across low-end ($ and $$ on Yelp, indicating $\leq$\$30 per person) and high-end ($$$ and $$$$ on Yelp; $>$\$30 per person) restaurants and across chain and independent restaurants respectively. On average, high-end restaurant reviews have more attributes (less *missing*) except for *location* which is mentioned more in low-end restaurant reviews. The ratings are generally more positive for high-end restaurants. Chain reviews tend to talk more about *service* and *location*, while

more experienced and have more friends.

[8]We identify restaurants as chains if they have multiple stores by the same name owned by a single firm.

reviews of independents talk about other attributes. The reviews for chains generally get an average of (3-star) or below on attributes whereas independents receive more 4 and 5-star attribute ratings.

Beyond the clear differences in attribute rating behavior (silence and valence) based on observables of restaurants and reviewers, the mapping from experience utilty to attribute ratings and attribute ratings to weights to obtain an overall rating can vary due to a variety of unobservables. To accommodate this, we allow for unobserved heterogeneity on these in our structural model.

1.6 Results

We describe the results in five parts: (i) attribute sentiment classification performance of various text mining methods; (ii) estimates of the structural model of rating behavior; (iii) drivers of attribute silence; (iv) validation; and (v) the impact of correcting for attribute silence.

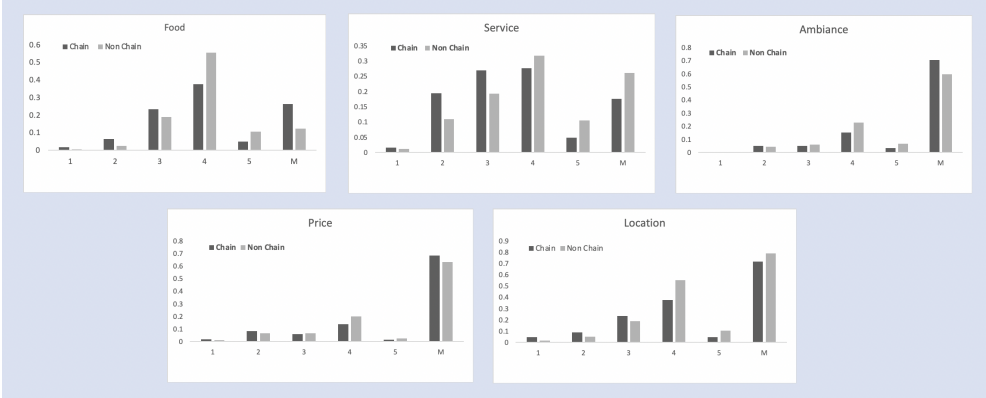
1.6.1 Attribute Sentiment Classification

We report the performance on attribute sentiment classification in three parts: (1) Overall classification accuracy; (2) Classification accuracy on "hard" sentene types; (3) polarity and attribute classification.

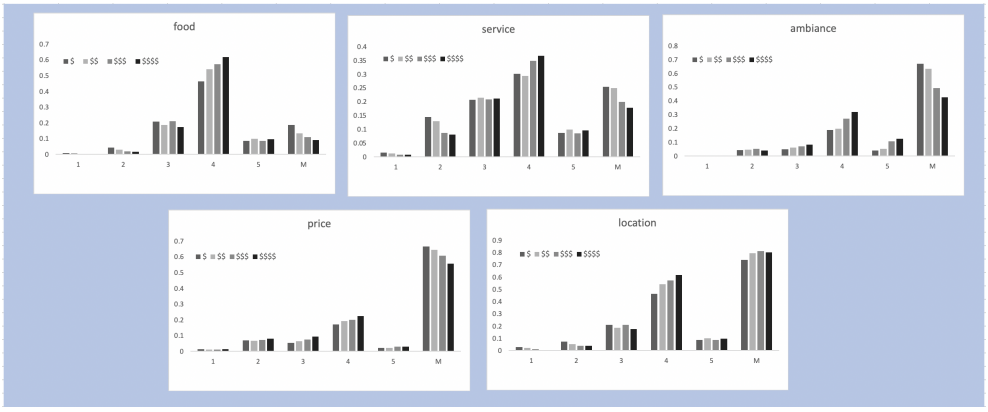
Overall Classification Accuracy. We begin by reporting the performance of the various models in terms of attribute and sentiment classification accuracy on the test dataset described earlier in the data section. The lexicon based method that relies on carefully crafted rules and human-tagged lexicons performs better than most supervised machine learning algorithms and is as good as the convolutional-LSTM in the attribute classification task. This is because this task is relatively unambiguous and the lexicons are constructed specific to the domain of restaurant reviews. However, this method does very poorly in the

Figure 1.4: Differences based on Observable Characteristics

(a) Elites and Non -Elites



(b) Chains and Independent Restaurants



(c) Price Ranges (1-4)

more complex 5-grained sentiment analysis task. Among supervised algorithms, Support Vector Machines (SVM) do better than most of the other classifiers in both attribute and sentiment classification tasks. This is in line with past literature that has shown that SVMs are the best Machine Learning based text classifiers. The network with only convolutional layer just matches the performance of the SVM. However, the convolutional-LSTM does better than all methods in both attribute and sentiment classification tasks. The accuracy of the convolutional-LSTM in the task of 5-level sentiment classification is 50%—lower than state of art accuracy 56% reported in (Brahma 2018), but on a different dataset for which we do not know the differential mix of "hard" versus "easy" sentences in the corpus. Further, they also do not provide metrics like confusion matrices, which helps assess other dimensions of classification accuracy.[9]

Table 1.9: Comparison of Text Mining Methods (I)

Type	Method	Attribute accuracy	Sentiment accuracy
Lexicon	Lexicon	68%	31%
Machine Learning	SVM	60%	40%
	Naives Bayes	43%	39%
	Logistic Regression	59%	41%
Deep Learning	CNN	62%	41%
	LSTM	62%	40%
	conv-LSTM (pre-trained)	**68**%	**47**%
	conv-LSTM (self-trained)	**71**%	**50**%

The convolutional-LSTM model with self-trained embeddings does slightly better than the one using pre-trained Glove embeddings both in terms of attribute and sentiment accuracy. This could be attributed to the slightly more relevant vocabulary generated when

[9]As an aside, we note that nlpprogress website which tracks state of the art (SOTA) for NLP tasks reports 72% accuracy using Yelp data as of 2019 for the 5 level sentiment task at the *review document* level. This is of course different from our 5 level *sentence level* sentiment task. But as a point of comparison, our model's performance for this document level task is 70%— comparable to the previous SOTA paper from 2017 (e.g., Johnson and Zhang 2017). Interestingly, we use a much smaller training data to achieve the same accuracy. While we make no claims in terms of being state of the art in terms of accuracy, we note that our classification results are in the ball park of "good" models. Our focus is on the performance on "hard" sentences.

Table 1.10: Comparison of Text Mining Methods (II)

Type	Method	Building Effort	Scalability	Interpretability
Lexicon	Lexicon	High	Low	High
Machine Learning	SVM Naives Bayes Logistic Regression	Moderate	High	Low
Deep Learning	CNN LSTM conv-LSTM (pre-trained) conv-LSTM (self-trained)	Moderate	High	Low

word vectors are trained from scratch on a specific corpus (For corpus differences, see the appendix at the end).

Classification Accuracy on Hard sentence types. To develop some intuition behind what drives the performance accuracy of these models, we test these models on simple and various types of hard sentences. We sampled 100 sentences of each type from the test dataset. Table 1.11 reports the comparative performance of the deep learning models, the best supervised machine learning model (SVM) and the lexicon method. As expected, the hybrid convolutional-LSTM performs better than most other models in all of these tough classification scenarios and especially in classifying scattered sentiment in long sentences. Interestingly, the convolutional-LSTM model does significantly better on simple sentences as well.

Table 1.11: Performance on Hard Sentence Types

	Simple	Hard(Overall)	Scattered	Implied	Contrastive
Lexicon	46%	17%	17%	18%	16%
SVM	47%	19%	18%	20%	20%
CNN	44%	21%	22%	17%	24%
LSTM	46%	30%	37%	28%	25%
Convolutional-LSTM	52%	34%	41%	31%	28%

Polarity and Attribute Classification. As we mentioned in the section *Performance Measures*, though accuracy is a first-order metric for hard problems like granular senti-

ment detection, we need other measures to refine model choice; especially among models with similar accuracy scores. Table 1.12 shows that the convolutional-LSTM model using Glove pre-trained embedding is slightly better than the one using self-trained embedding (though the overall accuracy is higher for the latter) because it preserves polarity better i.e. it mostly mis-classifies within the granular sentiment classes (positive, negative, neutral) and thus has lower polarity reversal.

Table 1.13 assesses attribute classification accuracy. We find that both the convolutional-LSTM based attribute classifiers using GloVe and self-trained embeddings do a fairly good job in classifying attributes across classes. Further, their performance is not driven simply by getting high-frequency classes like food right.

Table 1.12: Polarity Reversal Confusion Matrix (Sentiment Analysis)

	CNN			Convolutional-LSTM (self trained)			Convolutional-LSTM(Glove 300)		
True Class	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive
Very Negative	**31%**	7%	51%	**47%**	6%	47%	**71%**	2%	24%
Negative	**33%**	11%	63%	**45%**	6%	49%	**64%**	4%	35%
Neutral	14%	37%	49%	16%	33%	51%	44%	18%	39%
Positive	14%	9%	**77%**	15%	6%	**80%**	31%	5%	**64%**
Very Positive	9%	10%	**81%**	21%	2%	**88%**	18%	5%	**77%**

Table 1.13: Simple Confusion Matrix (Attribute Analysis)

	Convolutional-LSTM (self-trained)					Convolutional-LSTM (Glove 100)				
Predicted \ True	food	service	ambiance	value	location	food	service	ambiance	value	location
Food	**79%**	4%	2%	3%	2%	**75%**	6%	6%	3%	1%
Service	10%	**60%**	9%	5%	3%	7%	**76%**	8%	2%	0
Ambiance	8%	0	**58%**	3%	10%	2%	3%	**77%**	2%	2%
Value	10%	2	6%	**75%**	2%	8%	8%	6%	**74%**	4%
Location	8%	6%	11%	3%	**56%**	6%	14%	36%	3%	31%

1.6.2 Structural Model Estimates

Overall, we find a three segment model fits best.[10] Segment 1 the smallest segment, constitutes about 9% of the market. Segment 2, the largest segment accounts for 59% of the market, while Segment 3 constitutes 32% of the market.

[10] We assessed fit based on BIC for two, three and four segment models.

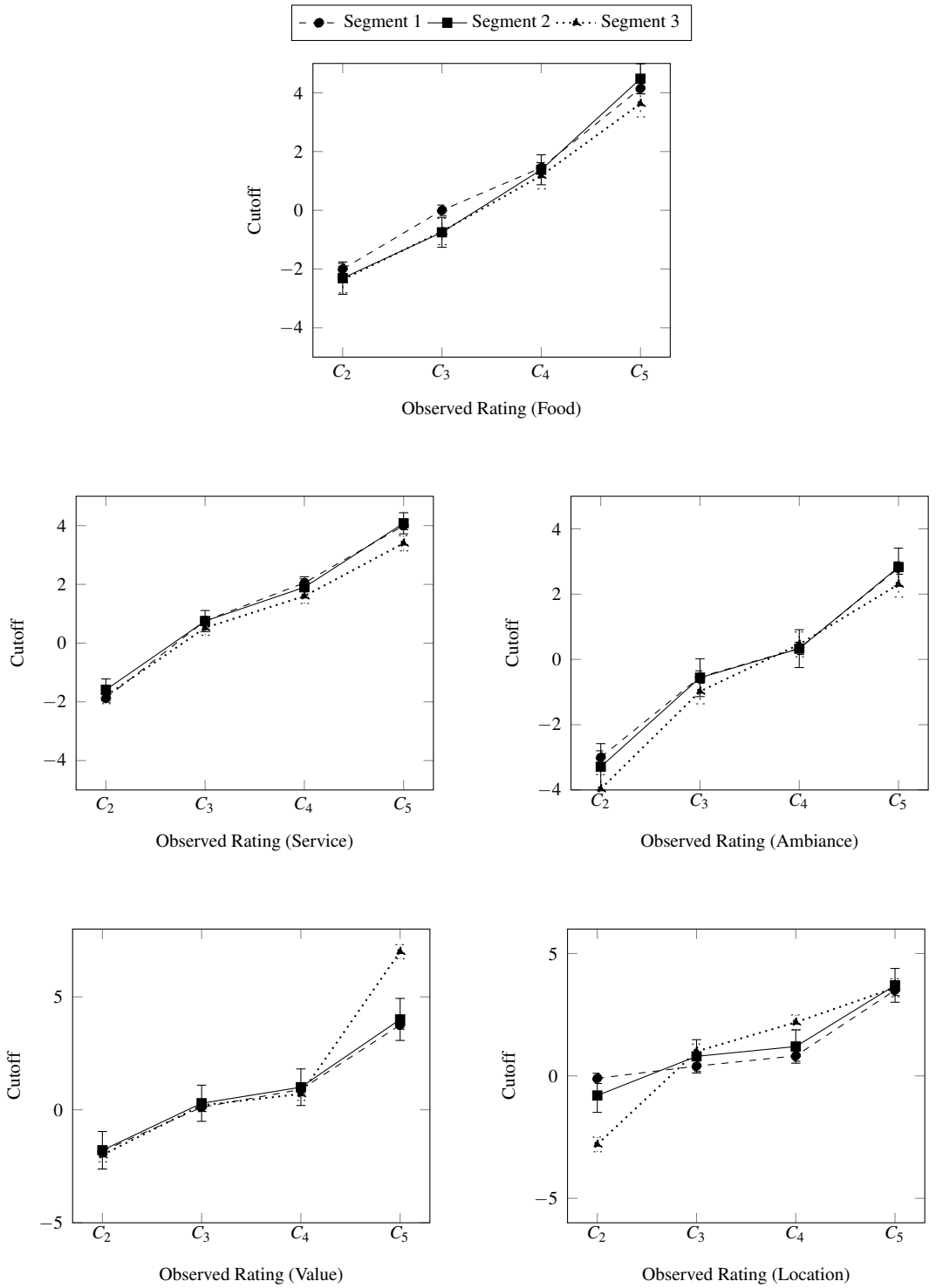
Ordinal Logit Model. The estimates of the ordinal logit model that maps latent utility to attribute ratings is presented in two parts. Table 1.14 presents the mapping between restaurant observables and true latent attribute level experience. As expected, restaurants with higher ratings have overall higher latent utility, chains have lower latent utility, and prices reduce latent utility. The thresholds $C_s(s \in 2,3,4,5)$— the cutoff between score $s-1$ and s of the ordinal logit for each of the three latent segments are shown in Figure 1.6. As expected, these thresholds are monotone and increasing in rating scale, but nonlinear. Getting higher score requires higher-quality experience across attributes as expected, but the marginal satisfaction required for each score is different across attributes, scores and reviewer segments. It should be noted that even though the thresholds often appear parallel, its implications for probability of a given rating for a segment is highly nonlinear and therefore heterogeneous. This is because there is much higher density in the middle than at the extremes.

Table 1.14: Structural Model Estimates
Link between Restaurant characteristics and attribute latent utility

	food	service	ambiance	value	location
Biz price $$	0.171*	0.107**	0.002 ***	0.036***	0.035***
	(0.093)	(0.051)	(0.002)	(0.013)	(0.016)
Biz price $$$	-0.007	0.306***	0.255***	-0.115***	0.356
	(0.084)	(0.0779)	(0.041)	(0.030)	(0.065)
Biz price $$$$	0.100**	0.269**	0.338***	-0.219***	0.650***
	(0.046)	(0.109)	(0.053)	(0.059)	(0.085)
Biz chain	-0.388***	-0.013	-0.079 **	-0.225 ***	-0.351***
	(0.089)	(0.032)	(0.050)	(0.036)	(0.041)
Biz average stars	0.263***	0.395***	0.249***	0.317***	0.254***
	(0.051)	(0.038)	(0.038)	(0.018)	(0.017)
Previous reviews (Star Rating)	0.338***	0.166***	0.046*	0.053***	0.027***
	(0.063)	(0.036)	(0.031)	(0.016)	(0.014)
N	38630	34636	17305	16227	10463

Overall Rating Regression. The weights on the attribute ratings that impact overall rating for the three latent segments are presented in Table 1.15. Note for ease of interpre-

Figure 1.5: Structural Model Estimates: Attribute Level Thresholds of Latent Utility by Segment



tation, the weights reported have been normalized such as the sum of the weights add to 1. Also, note that the model was estimated without normalization and all coefficients were estimated as positive.

Table 1.15: Structural Model Estimates
Attribute weights (normalized to sum to 1) on Ratings

	Segment 1	Segment 2	Segment 3
Food	0.229	0.322	0.01
Service	0.217	0.173	0.00
Ambiance	0.154	0.180	0.01
Value	0.194	0.163	0.39
Location	0.206	0.162	0.59
Segment size (by Review)	18%	58%	24%
Segment size (by Reviewer)	9%	59%	32%

Segment 1, the smallest at 9%, places the most importance on food in terms of their overall ratings. Segment 2, the largest at 59% cares not only about food, but also service. In contrast, for Segment 3, with 32% of reviewers, ratings are driven mostly about price and location.

Segment Interpretation. Finally, we report the descriptive statistics of each segment in Table 1.16 to aid interpretation. The smallest segment 1 (9% of reviewers) consists of 65% elites, writes most often and contributes double their share in reviews (18%). They write the longest reviews, and include the most number of attributes. They tend to write earlier than others on average. They tend to be harsher than the average rating of the restaurants and have relatively low variance of ratings. Given the high percentage of elites, greater frequency, and more comprehensive and longer reviews, we name them as "*status-seeking regulars*."

In contrast, Segment 3 accounting for 32% of reviewers has no elites, writes least frequently, contributing only 24% of reviews. The reviewers write the shortest reviews and include the fewest number of attributes. They tend to write at later stages after others have provided their reviews. They generally tend to be more generous in their overall

ratings. Interestingly, they also have the highest variance in their reviews, though they visit restaurants with high ratings and lower variance. We call them the ‘‘*emotive irregulars*,” given their lower frequency, and limited contributions in text reviews. They tend to offer either very positive or relatively negative reviews.

Finally, the largest segment 2 with 59% of the reviewers has only 26% elites. The reviewers are in the middle between Segment 1 and 3 in rating behaviors. They write fewer, shorter reviews and include fewer attributes than segment 1, but more than Segment 3. Their ratings are very similar to the average of the restaurant ratings. We call these reviewers as the “*altruistic mass*,” the bulk of the Yelp reviewing community, who write reviews diligently, but with little expectation of rewards or merely wanting their voice to be heard.

Table 1.16: Segment Characteristics

Characteristic	**Status-seeking Regulars** Mean (SD)	**Altruistic Mass** Mean (SD)	**Emotive Irregulars** Mean (SD)
% Elites	65%	26%	0%
Review Length (Chars)	889 (744)	677 (640)	349 (329)
No of Attributes	2.87 (1.1)	2.53 (1)	1.87 (0.83)
No of earlier reviews	22 (34.1)	24.2 (38)	36.7 (47.4)
Experience (Months)	33.6 (25.7)	24.6 (25.1)	16.9 (20)
Reviewer Rating	3.9 (0.4)	3.31 (1)	4.08 (1.2)
Business Rating	3.63 (0.7)	3.47 (1.1)	3.84 (0.7)
Proportion of Missing Attributes by Segment			
Food	0.10	0.18	0.22
Service	0.22	0.24	0.32
Ambiance	0.53	0.66	0.73
Value	0.58	0.62	0.89
Location	0.71	0.76	0.90

1.6.3 Drivers of Attribute Silence (Missingness)

With the estimates of the structural model, we now interpret attribute silence of each reviewer segment. We conjectured three plausible reasons driving attribute silence: (i) informativeness; (ii) attribute importance; and (iii) need to praise/vent. We assess each of these conjectures in turn.

Informativeness. Table 1.17 reports a logistic regression result with attribute presence as the DV. As conjectured, we expect experience attributes (food, service, ambiance) to be written more often than search attributes like price (which is a major component of value) and location. Further, if experience/search attribute is the driver of missingness, food and service should be missing more often at chains than at other restaurants. In addition to the attributes, the explanatory variables are positive and negative deviations in past attribute rating against predicted ratings; and the variance of the past attribute ratings.

The higher positive attribute coefficients (for food, service, ambiance) relative to the normalized location coefficient of zero, and value support our conjecture that reviewers write more often on experience attributes and tend to be more silent on search attributes which can be discovered easily on the site. Further, as expected, variance has a positive coefficient, supporting our hypothesis that attributes are more likely to be mentioned when opinions around that restaurant is not settled. Interestingly, for deviations, negative deviations induce the attribute to be mentioned, but vice versa for positive deviations. This is the case across all segments. Thus there is overall support for the informativeness conjecture. A subtle point from the results is that people are more likely to share information about unmet expectations (negative deviations) than positive deviations.

Attribute Importance. First, we compare the probability of missing attribute by segment in the bottom panel of Table 1.16 with the attribute importance weights of the three segments reported in Table 1.15. Food and service (and to a lesser extant ambiance) have the lowest rating of missing. Food, service and ambiance also have among the highest impact on overall ratings for Segments 1 and 2. But for segment 3, even though food and service do not drive overall ratings, they still are the most written about attributes. Similarly, even though value and location impact overall rating for Segment 3 these are still the most missing attributes in text reviews. Thus there is not a clear pattern that attribute missingness is driven by the importance of that attribute. To test this formally, we conducted a logistic regression with attribute presence as the DV and attribute importance as an ex-

Table 1.17: Impact of Informativeness on Attribute Presence (Segment-wise)

	Dependent variable:		
	Attribute Presence		
	(1)	(2)	(3)
Intercept	0.580^{***}	0.555^{***}	0.183^{***}
Food	0.153^{***}	0.169^{***}	0.273^{***}
Service	0.130^{***}	0.134^{***}	0.245^{***}
Ambiance	-0.035^{***}	0.017^{***}	0.063^{***}
Value	0.028^{***}	0.030^{***}	-0.020
Chain	0.104^{***}	0.078^{***}	0.079^{***}
Food $\times$ Chain	-0.129^{***}	-0.077^{***}	-0.187^{***}
Service $\times$ Chain	-0.065^{***}	-0.056^{***}	-0.103^{***}
Ambiance $\times$ Chain	-0.101^{***}	-0.112^{***}	-0.070^{*}
Value $\times$ Chain	-0.091^{***}	-0.067^{***}	0.027
Variance	0.061^{***}	0.059^{***}	0.068^{***}
Positive Difference	-0.378^{***}	-0.366^{***}	-0.241^{***}
Negative Difference	0.077^{***}	0.086^{***}	0.167^{***}
N	136,600	67,255	60,422
8,923			

Note: *p<0.1; **p<0.05; ***p<0.01

Positive Difference $= \|Actual\text{-}Own\|.I(Actual > Own)$,

Negative Difference $= \|Actual\text{-}Own\|.I(Actual < Own)$

planatory variable along with additional control variables. Interestingly, we do not find a significant positive effect on attribute importance. In fact, the regression results show a consistent negative effect. Thus our results question the conventional wisdom, and the implicit assumption underlying many topic models, that the frequency of occurrence of topics is implicitly assumed to be related to its importance. However, we temper our conclusion around importance, because food and service which are most present may also be the most important in driving the decision to visit a restaurant, but our estimated attribute importance is conditional on visit.

Table 1.18: Impact of Importance on Attribute Presence (Segment-wise)

	Dependent variable:		
	Attribute Presence		
	(1)	(2)	(3)
Intercept	0.597***	0.319***	0.317***
Food		0.612***	0.577***
Service		0.492***	0.448***
Ambiance		0.136***	0.146***
Value		0.122***	0.157***
Importance	−0.146***	−0.132***	−0.127***
Importance × Food			0.084***
Importance × Service			0.134***
Importance × Ambiance			−0.034
Importance × Value			−0.087***
Observations	148,605	148,605	148,605
Log Likelihood	−106,713.200	−87,654.920	−87,518.190
Akaike Inf. Crit.	213,430.400	175,321.800	175,056.400

Note: *p<0.1; **p<0.05; ***p<0.01

Praise/Vent Need. As discussed earlier, we report the missing odds as the ratio π^s_{gk} defined as $\frac{Pr(A_{gk}=s|M_{gk}=1)}{Pr(A_{gk}=s|M_{gk}=0)}$ for each attribute by segment as a function of predicted sentiment level in Figure 1.6. The patterns of attribute silence differ by attributes and by segment. For food, service and ambiance, all three segments tend to be more silent when they are dissatisfied, and write more when they are satisfied. However, segment 3 which places

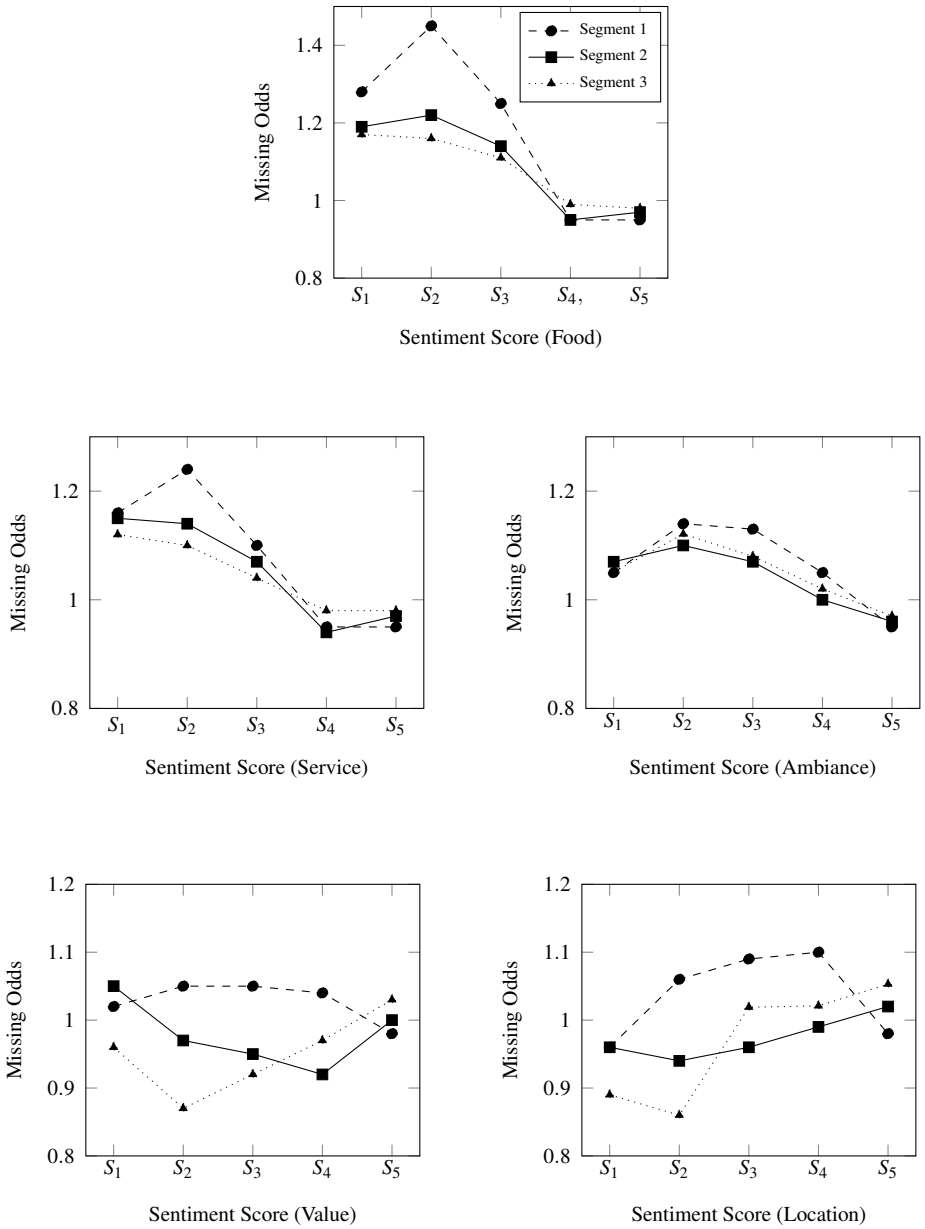
the most importance on location and value is more likely to write about these attributes when they are dissatisfied. This seems consistent with our label for them —as emotive irregulars. They don't write often, but they write when they are very satisfied with food, service and ambiance, but dissatisfied with value and location. This may also explain the higher variance in their overall ratings.

In summary, the information value of reviews play a significant role in the motivation to write about attributes across all segments. We also found the motivation to both praise good performance and vent about bad performance, but this varied across segments and attributes. For staple features like food, service and ambiance, all three segments are more likely to write when satisfied and less likely to write when dissatisfied. Overall, this might explain in general why reviews tend to be skewed to be more positive on rating sites—if this also translates to selection into who writes reviews. However segment 3 is likely to vent more when dissatisfied about two attributes that drive its ratings—value and location. The lack of a strong link between importance and mentions in reviews of attributes suggests that online reviews may not be as complete a source of topic and need identification as previously believed. However, we note that this could be because our attribute importance estimate are conditional on visit to restaurants, and may not account for its importance in decision to visit the restaurant. At the very least, our results suggest that we might want to be circumspect in the use of frequency of mentions as a proxy for benefit or need importance and explore this issue in future research.

1.6.4 Validation of Imputation

We validate our model-based imputation approach in Table 1.19 by assessing the ability to predict attribute ratings on a holdout sample, relative to no segmentation, where we assume reviewers have homogeneous rating styles, and ad-hoc imputation approaches, where reviewers who missed attribute ratings experienced average (score 3) or very low

Figure 1.6: Odds(π) of attribute missing in reviews as a function of sentiment level



(score 1) or very high (score 5) level of satisfaction. To be specific, we compare the predicted attribute ratings vs. observed rating if an attribute rating is present on hold-out sample (10% of the observations). The overall RMSE across all attributes is lower for our model relative to the benchmark models. Even when the RMSE is compared by attribute, we find that our model does better on all attributes, except location, where a uniform imputation of 3 can get slightly better prediction. Given the large share of missing data for location, the model identification was the weakest for this attribute. However, for all the other attributes the imputation from the model indeed does better.

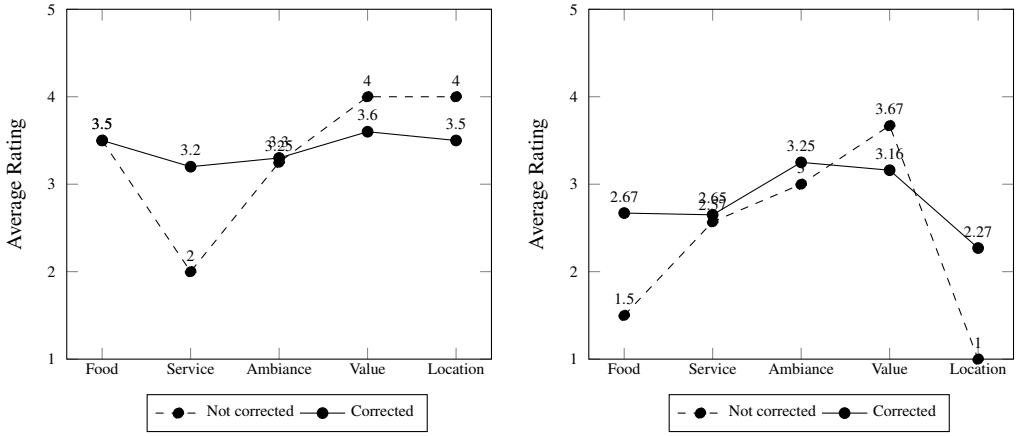
Table 1.19: Model Fit: Root Mean Squared Error (RMSE) across Imputations

Attribute	Our method	No Heterogeneity	Fixed Imputation Scores		
			Score 1	Score 3	Score 5
Overall	0.689	0.932	1.879	0.778	1.211
Food	0.699	0.881	2.631	0.973	1.323
Service	0.891	0.967	2.304	0.939	1.587
Ambiance	0.578	1.169	1.768	0.711	0.960
Value	0.664	0.924	1.534	0.706	1.132
Location	0.613	0.717	1.161	0.564	1.055

1.6.5 Correction for Attribute Silence in Attribute Ratings

We illustrate how correcting for attribute silence through imputation at the individual review level can impact overall attribute rating for a restaurant. We see that correction for missing attributes has significant impact on attributes that are missing more frequently: value and location in general, and food for chain restaurants. The correction could be either upward or downward depending on attribute, restaurant type and reviewer type. For example, at an independent restaurant in Phoenix where most reviewers are found to remain silent about service at higher satisfaction levels, observed service ratings are lower than actual service ratings after imputing for missing attribute ratings. Then, correction results in higher service ratings than observed ratings (Figure 1.7a). Food and ambiance

Figure 1.7: Change in Average Attribute Rating



(a) Independent restaurant in Phoenix (b) Chain restaurant in Las Vegas

scores barely change, and value and location scores slightly go up after imputation for this restaurant. In Figure 1.7b, we illustrate a chain restaurant in Las Vegas where many of the reviewers miss food and location attributes, when satisfied, and miss value rating when dissatisfied. Here food and location scores to go up and value score to go down. Overall this shows that our imputation approach based on restaurant observables, rater observable and unobservable heterogeneity is extremely flexible in its imputations and the ability to correct for missing attribute ratings.

Table 1.20: Impact of Imputation on Attribute Ratings

	Average Correction		% of corrections ≥ 0.5	
	Chain	Independent	Chain	Independent
Food	0.33	0.12	22%	1%
Service	0.24	0.32	13%	16%
Ambiance	0.83	0.62	83%	62%
Price	1.07	0.83	92%	91%
Location	1.29	1.16	92%	95%
N: 2719				

1.7 Conclusion

The paper addresses the general problem of using unstructured text data to generate quantifiable market feedback typically obtained through surveys; the specific application is to use restaurant reviews to generate attribute level ratings of restaurants. The paper addresses two novel and challenging problems around online text reviews: (i) convert text into *fine-grained numerical sentiment scores* on pre-specified attributes (e.g., food, service) by accounting for language structure; and (ii) accounting for attribute silence in attribute sentiment scoring. For the first problem, the it uses a deep learning convolution-LSTM model that exploits the spatial and sequential structure of language to improve sentiment classification, especially on known types of "hard" sentences in NLP. For addressing attribute silence, the paper develops and estimates a structural model of reviewer rating behavior that takes into account the data generating process to develop a model-based imputation procedure to address attribute silence. Overall, the paper illustrates the value of combining "engineering" thinking underlying machine learning approaches with "social science" thinking from econometrics to answer novel marketing questions.

Substantively, the paper identified three segments of reviewers—the smallest but most active reviewers ("Status Seeking Regulars,") the largest segment ("Altruistic Mass,") who review without reward expectations, and "Emotive Irregulars," who review infrequently, but write about attributes they are extremely satisfied or dissatisfied. Our insights around attribute silence in reviews shows that informativeness and need to praise/vent drive more of the writing than the importance of the attribute. Not only does this contribute to the literature on why people engage in online word of mouth (Berger 2014), it also has implications for using reviews as a source of data for needs/benefits identification. In particular, contrary to conventional wisdom, the frequency of mentions of a benefit or a topic may not necessarily be a proxy of its importance.

We conclude with a discussion of some suggestions for future research. First from the machine learning perspective, the research around improving performance on “hard” sentences needs to be pursued to further improve accuracy; while the performance improved for all of the hard sentence types there is more room for improvement. It would be useful to consider how recent methods such as BERT or GPT can improve on the fine-grained sentiment scoring problem for ‘’hard” sentences. From the substantive/econometric perspective, it would be useful to more systematically understand the drivers of attribute silence. While our current results offer suggestive evidence for our conjectures, a more systematic causal investigation of the attribute level motivations can further enrich the literature on the drivers of WOM. It would also be worth combining our content analysis at the attribute level with work on fake reviews/review shading to get a richer understanding of how to correct for these issues in tracking WOM.

Chapter 2

When do consumers talk?

2.1 Introduction

Many consumption decisions are influenced by what we learn from social connections, driving the explosion of user-generated information online. Indeed, empirical research shows that on average higher reviews tend to increase sales (Chevalier and Mayzlin (2006), Luca (2016), Liu et al. (2019)). This paper investigates, both theoretically and empirically, a strategic motive behind providing reviews, and explores how strategic communication drives the selection of user-generated content differentially, depending on the strength of the brand image.

We find a striking pattern for restaurant reviews on Yelp.com: On a 5-star scale, the modal rating is 1 star (46.9% in our data) for national established chain restaurants, but 4 or 5 stars for comparable independent restaurants (41.2%) in the same categories. Unless there are large systematic quality differences between chain and independent restaurants, this finding suggests positive or negative selection of content due to differences in the propensity to review after a positive versus a negative experience at these different types of restaurants. A selection effect has significant implications on how review data is interpreted by potential customers, and therefore on their purchase decisions and the firm's

revenues.[1] The goal of this paper is to shed some light on drivers of these selection effects.

We develop a model of word-of-mouth (WOM) communication that explains how positive or negative selection of WOM information arises in equilibrium. We identify two determining factors: strength of brand image, measured by the dispersion of consumer beliefs about product quality, and the informativeness of good and bad experiences.

Formally, we consider a monopolist who is launching a new product of uncertain quality. Some early adopters in the market get a chance to try the product and receive a private noisy binary signal of quality.[2] Then, the monopolist announces a net price for the product. In practice, the net price represents the price that potential consumers pay to purchase the product, which may be a combination of the posted price, promotions, extra benefits, etc. An early adopter can choose to share her product experience (signal) with a potential consumer, and influence his purchase decision. We characterize positive and negative WOM behavior in pure-strategy perfect Bayesian equilibria.

Our key premise is that writing reviews is costly, and early adopters share their experience only if they can instrumentally affect the purchase decision of the receiver of the message. This assumption is motivated by research in psychology and marketing that highlights two complementary functions of WOM: First, WOM helps consumers acquire information when they are uncertain about a purchase decision. Second, people engage in WOM to enhance their self-image, causing them to share information with instrumental value because this improves the image of the sharer as being smart or helpful.[3,4]

Given this assumption, the early adopter has to first take into account the probability with which the receiver of her message is also an early adopter — in which case WOM

[1] Reviews are well-known to be skewed (see Schoenmüller et al. (forthcoming)). Chevalier and Mayzlin (2006) and Fradkin et al. (2015) document positive skews in user ratings for books and home rentals, respectively.

[2] We do not model the purchase decision of early adopters in the main model, but discuss a possible dynamic extension when early adopters in the current period are followers from the previous period (Section 2.5.1).

[3] See Berger (2014) for a survey. The early adopter's incentive to share only instrumentally valuable information is also consistent with the persuasion motive of WOM, discussed in Berger (2014).

[4] Gilchrist and Sands (2016) instead consider WOM that brings pleasure in itself.

has no value — or is a consumer who has not tried the product yet (a follower). Next, in the case of a receiver who is a follower, what is important is his purchase decision in the absence of any WOM: This is determined by the price and the brand image defined by the distribution of followers' prior beliefs about quality. If a follower was likely to buy (not buy) given the brand image and the price only, then there is no reason for an early adopter to engage in positive (negative) WOM after a good (bad) experience, but she may affect the follower's action through negative (positive) WOM after a bad (good) experience.

The price set by the firm directly affects the follower's ex ante purchase decision, which indirectly affects WOM. For instance, by setting a high (low) price, followers are less (more) likely to buy ex ante, causing early adopters to engage in positive (negative) WOM. The strength of the brand image plays a critical role for how the monopolist sets the profit-maximizing price: If the brand image is well-entrenched, then all followers have the same identical beliefs about quality. So, the firm and early adopters can anticipate the followers' decision after receiving a message. But, if the brand is less-known or new, then followers don't know exactly what to expect resulting in heterogeneous beliefs about quality for idiosyncratic reasons. Then early adopters cannot predict the followers' decisions; some followers might buy after hearing positive WOM, while others might not buy despite positive news. This uncertainty crucially impacts the optimal pricing decision and the decision to engage in WOM in equilibrium.

First, we find that for well-entrenched brands, positive WOM cannot arise. If the fraction of new adopters is small, it is optimal for the firm to induce negative WOM. Intuitively, this is driven by the way "no WOM" is interpreted. If followers expect only negative experiences to be shared, then no WOM becomes a positive signal. With few early adopters, no WOM is observed with high probability, and so an equilibrium with negative WOM only is optimal for the firm. If the fraction of early adopters is above a threshold, then the number of early adopters with a negative signal increases, which decreases the benefit of a negative WOM equilibrium. In this case, the unique equilibrium involves no

WOM.

Second, we consider followers with heterogeneous beliefs. If the brand image is strong and close to well-entrenched, then only negative WOM can occur in equilibrium. If the brand image is weak, the type of WOM in equilibrium depends on the distribution of an early adopter's signal conditional on quality. We focus on equilibria when the fraction of new adopters is small. For the intuition, consider two extreme signal structures. If the signal structure is a "good news" process in that a positive experience is a strong signal for good quality, but a negative experience occurs with both good and bad quality, then the firm optimally sets a price that induces positive WOM. Conversely, for a "bad news" process, where a negative experience is very informativethe firm optimally induces only negative WOM.

Finally, using restaurant review data from Yelp.com and data on restaurant chains, we verify that our theory is consistent with empirical observation. We posit that consumers are likely to have homogeneous beliefs about restaurants that belong to a chain like Dunkin' with a strong brand image, but heterogeneous beliefs about independent restaurants like a new local coffee shop in New Haven. Controlling for restaurant characteristics (cuisine, price-range, location) and user characteristics (platform experience, average past ratings), our regression shows that being a chain restaurant results in approximately a 1-star reduction in every rating received relative to a similar independent restaurant. We also show that the propensity of a review being negative increases with the age of brand and the number of stores which can be thought of as proxies for brand strength. Our textual analysis of reviews further shows that reviewers are more likely to talk about prior beliefs (or *expectation*) when reviewing chain restaurants, especially in negative reviews, whereas they are more likely to anchor positive reviews of independent restaurants around the concept of *novelty*.

2.2 Literature Review

Our paper is substantively related to the research on diffusion of information through word-of-mouth, pioneered by Bass (1969). WOM can occur via platforms, social networks or traditional networks. Most early papers in this area treat WOM as a costless mechanical process, and focus on how the social network structure affects information percolation about the existence of a product: See for instance Galeotti (2010) or Galeotti and Goyal (2009).[5]

We contribute to the more recent literature that considers the strategic motive of consumers to engage in costly WOM. Campbell et al. (2017) focus on how the firm should balance WOM and advertising if consumers' incentive to talk stems from a desire to signal social status. They find that advertising crowds out consumers' incentives to engage in WOM. Other authors focus on WOM and referral programs. In Biyalogorsky et al. (2001) a firm can encourage WOM through the price or a referral program. Unlike in our model, a reduced price induces senders to talk because it "delights" them. Kornish and Li (2010) also consider the trade-off between referral rewards and pricing in a model where the sender cares about the receiver's surplus. Kamada and Öry (2017) consider a contracting problem in which the incentive to talk is driven by externalities of using a product together. They show that offering a free contract can make WOM more attractive since receivers are more likely to start using the product. We consider WOM not about the existence of a product, but about the experience. Early adopters engage in costly WOM only if their information has instrumental value and can affect the follower's action, and we characterize the connection between the firm's brand image and WOM.[6]

[5]Similarly, Leduc et al. (2017) study the diffusion of a new product when consumers learn about the quality in a network and the firm can affect the diffusion through pricing and referral incentives. Campbell (2013) instead analyzes the interaction of advertising and pricing. See also Godes et al. (2005) for a survey of the literature.

[6]The incentive to talk in our paper is similar to the incentive to search in Mayzlin and Shin (2011): The marginal value of information must be larger than the marginal cost of information dissemination or acquisition, respectively.

There is a growing empirical literature that studies the impact of review statistics, like volume, valence (positive or negative) and variance, on business outcomes (e.g., sales).[7] Luca (2016) finds that a one-star increase in Yelp ratings can decrease revenue by 5-9 percent. Chintagunta et al. (2010) show that an improvement in reviews leads to an increase in sales for movies and Seiler et al. (2017) documents that micro blogging has an impact on TV viewership. More specifically, the asymmetric impact of valence on profit-relevant outcome variables has been studied in some empirical contexts. Mittal et al. (1998) finds that negative information has larger impact on consumer purchase decisions compared to positive information. Chevalier and Mayzlin (2006) find that negative reviews have a larger effect on sales than positive reviews.[8] While the asymmetric impact of valence has been observed in several contexts, the literature does not explain what drives this asymmetry.

To the best of our knowledge, our paper is the first to provide an information-theoretical foundation for what determines valence of WOM and user-generated reviews, and to offer an explanation for the asymmetry observed empirically. We highlight how asymmetry in the propensity to engage in WOM can be driven by the dispersion of consumer beliefs about quality and the firm's pricing decision. The only other paper that studies different propensities to review after positive versus negative experiences is by Angelis et al. (2012), who argue using experimental evidence that consumers with a strong self-enhancement motive generate a lot of positive WOM, and transmit more negative WOM about other peoples' experiences: Differences in valence simply arise from differences in the type of people who choose to be early adopters. Chakraborty et al. (2019) also study selection in reviews using text analysis, but their focus is primarily on what drives content selection among different types of reviewers.

[7]For example, Nosko and Tadelis (2015), Dhar and Chang (2009) and Duan et al. (2008) show that the volume of reviews matter (rather than the rating), and Sun (2012) show that high variance in reviews corresponds to niche products, valued highly by some buyers but not by others. Onishi and Manchanda (2012) show a positive impact of blogging on sales.

[8]Also, Godes (2016) studies how the type of WOM affects the incentives of firms to invest in product quality.

We also contribute to the empirical literature on the relationship between branding and WOM. Luo (2009) finds that negative word of mouth has a medium-term and long-term effect on brand equity. Thus, even big established brands should be concerned about negative WOM and should try to understand how WOM evolves. Hollenbeck (2018) shows that value of franchising has declined with the rise of review platforms and thus small brands can now compete equally with larger brands. Unlike our paper, Hollenbeck (2018) does not address the issue of selection of reviews. His data indicates that differences in reviews can be broadly attributed to differences in quality, both for chain and non-chain hotels. However, since chain hotels systematically solicit WOM reviews from regular repeat customers, this may effectively eliminate potential negative selection.

2.3 Model

A firm produces a new product at a normalized marginal cost of zero. The quality $\theta \in \{H, L\}$ of the technology is high (H) with probability $\phi_0 \in [0,1]$, and is unknown to the firm.[9] The firm faces a continuum of consumers of measure 1. A fraction $\beta \in [0,1]$ of consumers are *early adopters* (he) who try the product first and observe a realized quality signal $q \in \{h, \ell\}$.[10] Given the type of technology θ, the realized quality q is drawn independently such that $Pr(q = h|\theta = H) = \pi_H$ and $Pr(q = h|\theta = L) = \pi_L$ where $1 \geq \pi_H > \pi_L \geq 0$. The remaining fraction $1 - \beta$ of consumers are called *followers* (she). Followers have not tried the product, and make their purchase decisions based on the expected quality.

Brand Image. It is useful to think of the consumer beliefs ϕ as reflecting the brand image. This is consistent with the standard interpretation, that consumer beliefs make up brand images which in turn influence consumer purchase decisions. For instance, Kotler (2000) writes: "A belief is a descriptive thought that a person holds about something. Beliefs

[9] Section 2.5.2 considers a privately informed firm.

[10] We can think of these early adopters as people who get introductory trial coupons or are invited for a soft launch of a restaurant.

may be based on knowledge, opinion, or faith (...) manufacturers are very interested in the beliefs that people have about their products and services. These beliefs make up product and brand images, and people act on their images."[11]

The distribution of these beliefs can therefore reflect the strength of the brand image – how consistent followers' knowledge is about ϕ_0. For products of a firm with a strong well-established brand image, it is reasonable to assume that consumers all know and agree on what to expect; in other words, followers observe ϕ_0. In that case, we say that beliefs are homogeneous and firm has a well-entrenched brand image. However, if the brand is less known or new, then followers are unlikely to know what ϕ_0 is, or in other words, followers can have different prior beliefs about the quality of the technology θ. To capture this idea, we assume consumers' priors ϕ are distributed according to a cdf F on $[0,1]$ with $\mathbb{E}^F[\phi] = \phi_0$ where F is independent of the actual quality. Formally, we analyze the following two cases separately:

- *Homogeneous priors:* All followers have the same prior belief ($F(\phi) = 1(\phi \geq \phi_0)$). This case reflects well-entrenched brands, where consumers know exactly what quality to expect. An example would be an outlet of a well-established chain like Dunkin'.

- *Heterogeneous priors:* Followers have idiosyncratic prior beliefs. We assume that F is continuous. This case will allow us to distinguish between strong and weak brand images based on the dispersion of buyer beliefs. See Section 2.4.2 for the formal definitions. To illustrate, a new independent coffee shop is likely to have a weak brand image and is therefore subject to dispersed idiosyncratic beliefs. Bigger chains, in turn, are likely to have more concentrated prior beliefs – in the extreme case being close to a well-entrenched brand image.

[11] Ke et al. (2020) model brand strength as dispersion of beliefs focusing on positioning rather than vertical quality.

Word-of-mouth Communication. Followers can potentially get information via word-of-mouth from early adopters. We assume that consumers are randomly matched in pairs. Thus, any consumer is matched to an early adopter with probability β. One can think of this as representative consumers who are most recently active on the review platform and want to leave a review for the next consumer who is visitng the platform, or individuals meeting off-line.[12] When consumers meet they do not know if they are matched to an early adopter or a follower.

Early adopters who have already consumed the product can obtain utility from sharing their signal with followers. We capture the incentives to engage in word-of-mouth with the following utility representation: Given his realized quality is $q \in \{h, \ell\}$, an early adopter's message space is $M_q := \{q, \emptyset\}$, i.e., communication is verifiable.[13] Engaging in WOM ($m = q$) entails a cost $c > 0$. An early adopter gets utility $r > 0$ if $q = h$ and the matched consumer buys, or if $q = \ell$ and the matched consumer does not buy. Put differently, the early adopter receives positive net utility from talking relative to not talking

1. either if $q = h$, he sends a message $m = h$, and the follower buys, but would not have bought with $m = \emptyset$ or

2. if $q = \ell$, he sends a message $m = \ell$, and the follower does not buys, but would have bought with $m = \emptyset$.

Our modeling is motivated by the self-enhancement and persuasion motives to talk for early adopters, and the information acquisition motive of followers, as described in Berger (2014). He argues that when people care about impression management, they are "more likely to share things that make them look good rather than bad." Importantly, the early adopter does not care about the ex-post quality realization of the follower. Instead he only cares about sending a message that is useful to the receiver in the interim for her purchase

[12] The case in which one review is read by more than one follower, is discussed in Section 2.5.3.

[13] We do not consider review manipulation as in Mayzlin et al. (2014), Luca and Zervas (2016), and He et al. (2020).

decision. So, talking can be effectively interpreted as the early adopter's impression management or self-enhancement and r is the early adopter's utility of an enhanced self-image from providing information of instrumental value.[14] Because messages are verifiable, the utility specification above reflects also the persuasive motive, where a sender engages in word-of-mouth to influence others and change their action. [15]

Timing and payoffs. The game proceeds as follows:

0. Some early adopters experience a quality realization q.

1. The firm chooses price p.

2. Early adopters decide whether to engage in WOM by sharing $m \in M_q$.

3. Each follower updates her belief about θ, and decides whether to buy or not.

We do not model how early adopters came to try the product in the first place, because we want to focus on the incentive to engage in WOM. In Section 2.5.1, we discuss how our baseline model can be extended to a dynamic setting in which today's followers become tomorrow's early adopters.

Histories, strategies, and equilibrium. A firm's strategy comprises a price $p \in [0,1]$. An early adopter's set of histories is $\mathscr{H}^{\mathrm{a}} = [0,1] \times \{h,\ell\}$ and his WOM strategy $\mu : \mathscr{H}^{\mathrm{a}} \to M := M_h \cup M_\ell$ maps the price and signal $q \in \{h,\ell\}$ to a message, where $supp(\mu(p,q)) = \{q, \emptyset\}$. A follower's history is in $\mathscr{H}^{\mathrm{f}} = [0,1] \times M \times [0,1]$ and her purchasing strategy $\alpha : \mathscr{H}^{\mathrm{f}} \to \{\text{buy}, \text{not buy}\}$ maps p, the message received $m \in M$ and her prior ϕ to a purchasing decision. We consider *perfect Bayesian equilibria (PBE) in pure strategies*. A PBE comprises a tuple $\{p, \mu, \alpha, \hat{\phi}\}$ such that all players play mutual best-responses given their beliefs about θ, where $\hat{\phi}(\phi, m)$ describe a follower's posterior belief given prior ϕ

[14] Restaurant reviewers on Yelp.com cite simplified decision-making for first-time visitors as one of the reasons for writing a review. See Carman (2018).

[15] This is also consistent with the Gricean maxims proposed in Grice et al. (1975) that when engaging in a conversation, people should make it relevant to the audience and provide enough information, but not more than required. We thank Kristin Diehl and Gizem Ceylan-Hopper for pointing us to this reference.

and message m. Let $\mu_q(p) \in \{0,1\}$ denote the probability with which an early adopter, who sees signal q and price p, engages in WOM in equilibrium. We omit p and write μ_q if there is no ambiguity.

Let $\xi := \frac{c}{r}$. We assume $1-\beta > \xi$, to rule out the trivial case of early adopters never engaging in WOM because they are unlikely to face a follower.

2.4 Equilibrium Characterization

We proceed by backwards induction and start with the sub-game after the price is set. We call this the "WOM subgame" and its equilibria "WOM equilibria." Proofs are in the Appendix.

2.4.1 Word-of-Mouth subgame

Purchase decision of a follower: It is optimal for a follower with prior ϕ and message m to purchase if and only if her expected utility from purchasing exceeds the outside option:

$$\hat{\phi}(\phi,m)\pi_H + (1-\hat{\phi}(\phi,m))\pi_L - p \geq 0.$$

Let $\Phi(p)$ denote the posterior belief that makes a follower indifferent between buying and not, i.e.,

$$\Phi(p) := \frac{p-\pi_L}{\pi_H-\pi_L}.$$

Then, a follower's best response is

$$\alpha(p,\phi,m) = \begin{cases} \text{buy} & \text{if } \hat{\phi}(\phi,m) > \Phi(p) \\ \text{buy or not buy} & \text{if } \hat{\phi}(\phi,m) = \Phi(p) \\ \text{not buy} & \text{otherwise} \end{cases}. \tag{α}$$

A follower's posterior belief after message $m \in \{h, \ell\}$ is $\hat{\phi}(\phi, h) = \frac{\phi \pi_H}{\phi \pi_H + (1-\phi)\pi_L}$ and $\hat{\phi}(\phi, \ell) = \frac{\phi(1-\pi_H)}{\phi(1-\pi_H)+(1-\phi)(1-\pi_L)}$, respectively. If the early adopter sends no WOM message ($m = \emptyset$), then the posterior depends on the equilibrium strategy of the early adopter captured by μ_h and μ_ℓ, so by Bayes' rule:

$$\hat{\phi}(\phi, \emptyset) = \frac{\phi \ [1 - \beta + \beta\,(\pi_H(1-\mu_h) + (1-\pi_H)(1-\mu_\ell))]}{1 - \beta + \phi\beta\,(\pi_H(1-\mu_h) + (1-\pi_H)(1-\mu_\ell)) + (1-\phi)\beta\,(\pi_L(1-\mu_h) + (1-\pi_L)(1-\mu_\ell))}.$$

Note that $\hat{\phi}(\phi, h) \geq \hat{\phi}(\phi, \emptyset) \geq \hat{\phi}(\phi, \ell)$, but $\hat{\phi}(\phi, \emptyset)$ can be higher or lower than the prior ϕ. The follower gets *"good news"* about the product if $\hat{\phi}(\phi, m) > \phi$ and *"bad news"* if $\hat{\phi}(\phi, m) < \phi$.[16]

Followers' posterior beliefs $\hat{\phi}(\phi, m)$ and strategy α define thresholds, such that after a message m, a follower purchases only if his prior is above this threshold. Let $\bar{\phi}(p)$ be such that after $m = \ell$, it is optimal to buy if and only if $\phi \geq \bar{\phi}(p)$, i.e.,

$$\bar{\phi}(p) = \frac{1}{\frac{1-\pi_H}{1-\pi_L}\,\frac{1-\Phi(p)}{\Phi(p)} + 1}.$$

Similarly, let $\underline{\phi}(p)$ be such that after $m = h$, it is optimal to buy if and only if $\phi \geq \underline{\phi}(p)$, i.e.,

$$\underline{\phi}(p) = \frac{1}{\frac{\pi_H}{\pi_L}\,\frac{1-\Phi(p)}{\Phi(p)} + 1}.$$

Finally, let $\tilde{\phi}(p; (\mu_h, \mu_\ell))$ be such that after $m = \emptyset$, it is optimal to buy if and only if $\phi \geq \tilde{\phi}(p; (\mu_h, \mu_\ell))$, i.e.,

$$\tilde{\phi}(p; (\mu_h(p), \mu_\ell(p))) = \frac{1}{1 + \frac{1-\beta+\beta(\mu_h(p)(1-\pi_H)+\mu_\ell(p)\pi_H)}{1-\beta+\beta(\mu_h(p)(1-\pi_L)+\mu_\ell(p)\pi_L)}\,\frac{1-\Phi(p)}{\Phi(p)}},$$

given message strategy $(\mu_h(p), \mu_\ell(p))$. Figure 2.1 summarizes these thresholds which characterize the follower's best response α. We call a WOM equilibrium

[16] $m \in \{h, \ell\}$ is verifiable and $m = \emptyset$ is on-path since a follower is matched to another follower with positive

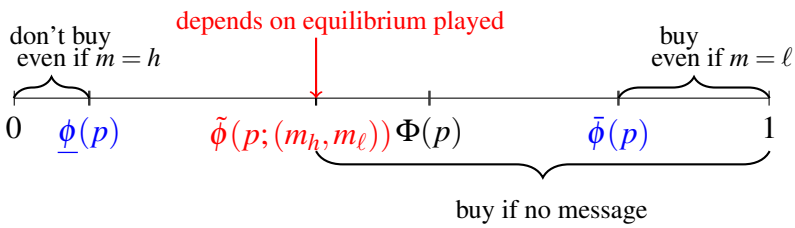


Figure 2.1: Followers' decisions for given prior beliefs

1. *full WOM equilibrium* if $\mu_h = \mu_\ell = 1$. Then, $\tilde{\phi}(p;(1,1)) = \Phi(p)$;

2. *no WOM equilibrium* if $\mu_h = \mu_\ell = 0$. Then, $\tilde{\phi}(p;(0,0)) = \Phi(p)$;

3. *negative WOM* if $\mu_h = 0, \mu_\ell = 1$. Then,

$$\Phi(p) \geq \tilde{\phi}(p;(0,1)) = \frac{1}{1 + \frac{1-\beta+\beta\pi_H}{1-\beta+\beta\pi_L}\frac{1-\Phi(p)}{\Phi(p)}};$$

4. *positive WOM* if $\mu_h = 1, \mu_\ell = 0$. Then

$$\Phi(p) \leq \tilde{\phi}(p;(1,0)) = \frac{1}{1 + \frac{1-\beta+\beta(1-\pi_H)}{1-\beta+\beta(1-\pi_L)}\frac{1-\Phi(p)}{\Phi(p)}}.$$

The absence of WOM ($m = \emptyset$) means "good news" in a negative WOM equilibrium, but "bad news" in a positive WOM equilibrium. The number of early adopters β determines the informativeness of $m = \emptyset$. It is a weaker signal, the less likely a follower is matched to an early adopter (β small).

Early adopter's WOM decision: Assume F has no mass point at the thresholds $\underline{\phi}(p), \overline{\phi}(p), \tilde{\phi}(p;(\mu_h,\mu_\ell))$ for $\mu_h, \mu_\ell \in \{0,1\}$. Then, an early adopter who observes $q = h$

probability.

weakly prefers to engage in WOM whenever

$$\underbrace{(1-\beta)r\big(F(\tilde{\phi}(p;(\mu_h,\mu_\ell)))-F(\underline{\phi}(p))\big)}_{\text{benefit of talking if } q=h} \geq \underbrace{c}_{\text{cost of talking}} .$$

Similarly, if $q=\ell$, an early adopter weakly prefers to engage in WOM whenever

$$\underbrace{(1-\beta)r\big(F(\bar{\phi}(p))-F(\tilde{\phi}(p;(\mu_h,\mu_\ell))\big)}_{\text{benefit of talking if } q=\ell} \geq \underbrace{c}_{\text{cost of talking}} .$$

To characterize the WOM equilibrium, we call followers

- *pessimistic*, whenever $F(\Phi(p))-F(\underline{\phi}(p)) \geq \frac{\xi}{1-\beta} \geq F(\bar{\phi}(p))-F(\Phi(p))$;
- *optimistic*, whenever $F(\bar{\phi}(p))-F(\Phi(p)) \geq \frac{\xi}{1-\beta} \geq F(\Phi(p))-F(\underline{\phi}(p))$;
- *uninformed* whenever $F(\Phi(p))-F(\underline{\phi}(p)),\ F(\bar{\phi}(p))-F(\Phi(p)) \geq \frac{\xi}{1-\beta}$;
- *well-informed* whenever $F(\Phi(p))-F(\underline{\phi}(p)),\ F(\bar{\phi}(p))-F(\Phi(p)) \leq \frac{\xi}{1-\beta}$.

Importantly, this definition is independent of the WOM equilibrium played.

[WOM sub-game] Let price p be such that F has no mass point at $\underline{\phi}(p)$, $\overline{\phi}(p)$, $\tilde{\phi}(p;(\mu_h,\mu_\ell))$ for $\mu_h,\mu_\ell \in \{0,1\}$. There exist thresholds $\hat{\beta}^{\text{neg}}(p),\hat{\beta}^{\text{pos}}(p)>0$ such that

1. A **full WOM equilibrium** exists if and only if followers are uninformed.
2. A **no WOM equilibrium** exists if and only if followers are well-informed.
3. A **negative WOM equilibrium** exists for all $\beta \in [0,1]$ if followers are optimistic. For $\beta < \hat{\beta}^{\text{neg}}(p)$, a negative WOM equilibrium does not exist if buyers are not optimistic.
4. A **positive WOM equilibrium** exists for all $\beta \in [0,1]$ if followers are pessimistic. For $\beta < \hat{\beta}^{\text{pos}}(p)$, a positive WOM equilibrium does not exist if buyers are not pessimistic.

An early adopter incurs the WOM cost only if the followers' decision is affected with a sufficiently high probability. Thus, with pessimistic followers, early adopters with a positive experience have a strong incentive to talk, while those with a negative experience have a weaker incentive. Indeed, in that case, a positive WOM equilibrium exists and $m = \emptyset$ is bad news. Similarly, with optimistic priors a negative WOM equilibrium exists. With well-informed followers, a large proportion of followers cannot be influenced, implying that there is no WOM. Analogously, with uninformed followers, the unique WOM equilibrium entails full WOM.

Multiplicity arises for large β. For example, with pessimistic followers, in a positive WOM equilibrium, $m = \emptyset$ is bad news and is almost equivalent to $m = \ell$. Thus, a negative WOM equilibrium also exists. The case when F has mass-points at the thresholds, is considered in the proof of Proposition 2.4.2.

2.4.2 Main results

Finally, we consider the full game including the firm's pricing decision. Define $\pi(\phi_0) := \phi_0\pi_H + (1-\phi_0)\pi_L$ to be the firm's belief that an early adopter has a good experience.

Homogeneous priors

[Homogeneous priors or well-entrenched brand image] Let $F = 1(\phi \geq \phi_0)$. In any pure-strategy equilibrium, negative WOM can be sustained in equilibrium if and only if

$$\beta \leq \bar{\beta}^{\text{hom}} := \frac{(1-\phi_0)\phi_0(\pi_H - \pi_L)^2}{(1-\pi(\phi_0))(\pi(\phi_0) - (\phi_0\pi_H^2 + (1-\phi_0)\pi_L^2))}.$$

No WOM can be sustained if and only if $\beta \geq \bar{\beta}^{\text{hom}}$. No other WOM equilibria can be sustained.

Intuitively, for a well-entrenched brand (if all followers have the same belief), the firm can set a price low enough such that all followers buy in the absence of WOM. The firm

cannot improve upon this. For small β, the firm can increase the price if $m = \emptyset$ is a weak good signal, which is the case in a negative WOM equilibrium. Followers who receive a negative signal will not buy, but for small β, there are only few such followers. If β is large, negative WOM is not worthwhile because too many followers receive the negative signal. Positive WOM is worthwhile only if the firm can charge a higher price to followers with a positive message. However, this is dominated by no WOM, where all consumers buy.

Heterogeneous priors

Next, consider heterogeneous priors, with continuous F. Denote the set of profit-maximizing prices by

$$\mathscr{P}^* = \arg\max_{p \in [0,1]} p(1 - F(\Phi(p))).$$

Note that $\mathscr{P}^* \neq \emptyset$ because the prices in the set are maximizing a continuous function on a compact set. We first focus on two cases:

1. Strong brand image: For all $p \in \mathscr{P}^*$, $F(\Phi(p)) < \xi$.

2. Weak brand image: For all $p \in \mathscr{P}^*$, $F(\Phi(p)) > \xi$.

For a product with a strong brand image, consumers have relatively concentrated beliefs so that any static profit-maximizing price can incentivize most buyers to buy. Put differently, there are not many buyers who can be convinced to buy as a result of receiving positive WOM. Note that the homogeneous prior (or well-entrenched brand image) case is the limit of strong brand image distributions, and indeed no positive WOM can occur. [Heterogeneous priors: Strong and weak brand image]

1. If the firm has a strong brand image, then for sufficiently small β, any pure strategy equilibrium entails no positive WOM.

2. If the firm has a weak brand image and $\pi_L = 0$, then for sufficiently small β, any equilibrium entails positive WOM.

If the profit maximizing price is unique, i.e. $\mathscr{P}^*$ is a singleton, then there is a unique (generically in ξ) equilibrium. In this case, we can fully characterize the equilibrium for sufficiently small β. The characterization also highlights the role of the signal structure induced by the early adopter's experience. [WOM under heterogeneous priors] Consider $\mathscr{P}^* = \{p^*\}$. For thresholds $\overline{\xi} := F(\overline{\phi}(p^*) - F(\Phi(p^*)), \underline{\xi} := F(\Phi(p^*)) - F(\underline{\phi}(p^*) > 0$ and sufficiently small β, there is a generically unique pure strategy equilibrium with the profit maximizing price being close to p^*. If $\xi < \min\{\underline{\xi}, \overline{\xi}\}$, it entails full WOM. If $\xi > \max\{\underline{\xi}, \overline{\xi}\}$, no WOM arises. If $\overline{\xi} > \underline{\xi}$ and for $\xi \in (\underline{\xi}, \overline{\xi})$, it entails negative WOM. If $\overline{\xi} < \underline{\xi}$ and $\xi \in (\overline{\xi}, \underline{\xi})$, it entails positive WOM. Intuitively, as $\beta \to 0$, under any WOM regime, the demand converges uniformly to $1 - F(\Phi(p)$. Hence, the profit maximizing price in any equilibrium converges to p^*. The type of WOM is determined by where ξ lies, relative to $F(\Phi(p^*)) - F(\underline{\phi}(p^*)$ and $F(\overline{\phi}(p^*) - F(\Phi(p^*))$. With heterogeneous priors, we focus on small β due to the richness of equilibria, and because WOM is most relevant for new products where the number of adopters is still small.

To understand the role of the signal structure, consider the example of $F = U[0,1]$. Then, $\overline{\xi} = \frac{\pi_H(\pi_H - 2\pi_L)}{2(\pi_H - \pi_L)(2 - \pi_H - 2\pi_L)}$ and $\underline{\xi} = \frac{\pi_H(\pi_H - 2\pi_L)}{2(\pi_H - \pi_L)(\pi_H + 2\pi_L)}$, so $\overline{\xi} > \underline{\xi} \Leftrightarrow 2\pi_L > 1 - \pi_H$. Think about two limiting cases. Suppose $\pi_L \approx 0$, i.e., it is unlikely for a bad firm to be able to generate a good experience. Hence, $q = h$ is particularly informative since it fully reveals that $\theta = H$: Examples are categories like independent restaurants where the consumer is discerning and is looking for specialized qualities. In such situations, negative WOM is never optimally induced by the firm, but positive WOM is induced for an intermediate range of WOM costs. For $\pi_L = 0$ we have $\overline{\xi} = \frac{\pi_H}{2(2-\pi_H)} < \underline{\xi} = \frac{1}{2}$ and $\overline{\xi}$ is increasing in π_H. Thus, positive WOM is optimal for a wider range of costs if π_H is also small. Next, suppose $\pi_H \approx 1$. Then, a good firm can generate a positive experience with high likelihood.

Here, $q = l$ is particularly informative. Car rentals might fall into this category: Customers are happy as long as no major quality flaws such as cleanliness or terrible service occur. In this case, $\overline{\xi} = \frac{1}{2(1-\pi_L)} > \underline{\xi} = \frac{1-2\pi_L}{2(1+\pi_L-2\pi_L^2)}$ and $\overline{\xi} - \underline{\xi}$ is increasing in π_L. Now, negative WOM is optimal for a wider range of costs if π_L is large. Finally, if $\pi_H < 2\pi_L$, then the firm induces no WOM because no signal is sufficiently informative about quality.

2.5 Extensions

Our baseline model is kept as lean as possible to highlight our main results, that we then test in the data. In this section, we discuss how far our results generalize in various dimensions.

2.5.1 Dynamics

We consider a single round of WOM decisions followed by purchasing decisions. A natural extension is to allow for dynamics, where followers today may in turn engage in WOM tomorrow. Consider the following alternative model. As before, there is a unit mass of consumers, and the prior belief about the firm's unknown product quality, at the start of the game, is given by $\phi^0 = \phi \sim F$. Time $t = 0, 1, 2, \ldots$ is discrete. At $t = 0$, a fraction β^0 of early adopters tries the product for free. In every subsequent period t, early adopters comprise the early adopters from period 0 and all followers who have adopted up to period t. The timing of the dynamic game is the natural analog of the static game of the baseline model, and proceeds as follows:

1. In every period t, a consumer is randomly chosen to be a potential reviewer (engage in WOM), and a fraction $\Delta > 0$ of consumers is picked at random to be potential followers.

2. The firm sets a period-t net price p_t.

3. If the chosen potential reviewer is not an early adopter, then there is no WOM in period t. If he is an early adopter, then he has experienced a quality signal q in some period prior to t, and can decide whether to share it or not. Payoffs of early adopters in any period t are analogous to those in the baseline model, i.e., he receives a benefit $\frac{r}{\Delta}$ from every follower who adopts in that period. Formally, he engages in WOM after experiencing $q = h$ whenever

$$(1-\beta^t)r(F(\tilde{\phi}(p_t;(\mu_h(p_t),\mu_\ell(p_t)))) - F(\underline{\phi}(p_t))) \geq c$$

and engages in WOM after experiencing $q = \ell$ whenever

$$(1-\beta^t)r(F(\overline{\phi}(p_t)) - F(\tilde{\phi}(p_t;(\mu_h(p_t),\mu_\ell(p_t))))) \geq c,$$

4. Finally, potential followers in period t decide whether to buy or not based, on their updated belief about quality. Again, analogous to our baseline model, this belief of a follower in period t can be calculated by Bayes' rule, using the consumer belief from period $t-1$ and the WOM message (or lack of WOM) in period t.

In this setting, both the distribution of priors ϕ^t and the fraction of early adopters β^t are changing over time, but the equilibrium outcome in each period is derived exactly as in our baseline model. Thus, the results in Proposition 2.4.2 generalize. In equilibrium, negative WOM arises early on, followed by no WOM later when β^t exceeds a threshold. The results in Propositions 2.4.2 and 2.4.2 are valid in periods in which β^t is sufficiently small, given the distribution of period-t priors ϕ^t.

2.5.2 Private firm type

In the baseline model, we assumed that the firm was unaware of its quality when it set its price. This captures situations where the firm is launching an entirely new product and

does not know about product efficacy prior to a large-scale launch. However, in other settings, the firm may be aware of its quality at the time of its pricing decision. In this section, we consider a straightforward extension, now with private information. For simplicity, we focus on situations with few early adopters (small β), and the uniform distribution in the case of heterogeneous priors.

If a firm has private information about quality, then it can, in principle, signal this information through its price, and followers may update their beliefs about the firm's type. However, such signaling via prices cannot arise in a pure-strategy equilibrium, i.e., there is no fully separating equilibrium.[17] To see this, note that in a fully separating equilibrium, all buyers are willing to buy at any price $p \leq \pi_H$. Thus, if $p_H > p_L$, then the L-firm wishes to deviate to offering p_H. If $p_L > p_H \geq \pi_L$, then no one buys at the price p_L and the L-firm can increase profits by deviating to p_H. Consequently, any pure-strategy equilibrium must be pooling, that is both firm types choose the same price. In such an equilibrium, the posterior belief is independent of the observed price. We characterize the unique pooling equilibrium in which the H-type firm maximizes its profits. This equilibrium has similar features to the equilibrium constructed in Section 2.4. In particular, the WOM subgame is identical and Lemma 2.4.1 applies. However, the profit function differs, as a θ-type firm can now calibrate demand using its private information about its quality. The following proposition is the analog to the results in the baseline model.

1. Consider a setting with homogeneous priors. For sufficiently small β, firms induce a negative WOM equilibrium in any pooling equilibrium.

2. Consider a setting with heterogeneous priors and $F = U[0,1]$. For sufficiently small β, in the H-optimal pooling equilibrium, given the same cutoff costs $\overline{\xi}$ and $\underline{\xi}$ as Proposition 2.4.2, the equilibrium entails full WOM if $\xi \leq \min\{\underline{\xi}, \overline{\xi}\}$ and no WOM if $\xi \geq \max\{\underline{\xi}, \overline{\xi}\}$. Furthermore,

[17] We conjecture that a semi-separating may exist if we allowed for mixed strategies.

- if $2\pi_L \geq 1-\pi_H$, then $\overline{\xi} \geq \underline{\xi}$ and for $\xi \in [\underline{\xi}, \overline{\xi}]$ the equilibrium entails negative WOM,
- if $2\pi_L \leq 1-\pi_H$, then $\overline{\xi} \leq \underline{\xi}$ and for $\xi \in [\overline{\xi}, \underline{\xi}]$ the equilibrium entails positive WOM.

To summarize, only the profit-maximizing price differs from the setting with symmetric information. All WOM equilibria are unchanged.

2.5.3 More than one follower

In a platform like Yelp, a single review is read by multiple potential consumers. In our model, if an early adopter is matched to not one, but $n > 1$ followers whose prior beliefs are drawn from a distribution F, the analysis is identical, but with ξ replaced by $\frac{\xi}{n}$. Thus, the more followers can see a review, the more word of mouth we expect. However, the type of word of mouth is unaffected by the number of followers n that one receiver speaks to. Note that, in this argument, we do not take into account that followers might be receiving more than one message. Instead, an early adopter is chosen at random to have the opportunity to review and n followers read the review.

2.5.4 Idiosyncratic value

Finally, one might wonder how the results would change if the potential customer base had idiosyncratic preferences over different products (horizontal differentiation). Suppose that the expected utility of a follower of purchasing the product at price p is

$$\hat{\phi}(\phi, m)\pi_H + (1-\hat{\phi}(\phi, m))\pi_L + \varepsilon - p \geq 0$$

where ε is a taste parameter distributed according to a distribution G. The cutoff beliefs $\Phi(p, \varepsilon)$, $\overline{\phi}(p, \varepsilon)$ and $\underline{\phi}(p, \varepsilon)$ are functions of the realized ε and we need to categorize

WOM equilibria as

- *pessimistic*, whenever $\int F(\Phi(p,\varepsilon)) - F(\underline{\phi}(p,\varepsilon))dG(\varepsilon) \geq \frac{\xi}{1-\beta} \geq \int F(\bar{\phi}(p,\varepsilon)) - F(\Phi(p,\varepsilon))\,dG(\varepsilon)$;

- *optimistic*, whenever $\int F(\bar{\phi}(p,\varepsilon)) - F(\Phi(p,\varepsilon))\,dG(\varepsilon) \geq \frac{\xi}{1-\beta} \geq \int F(\Phi(p,\varepsilon)) - F(\underline{\phi}(p,\varepsilon))dG(\varepsilon)$;

- *uninformed* whenever $\int F(\Phi(p,\varepsilon)) - F(\underline{\phi}(p,\varepsilon))dG(\varepsilon)$, $\int F(\bar{\phi}(p,\varepsilon)) - F(\Phi(p,\varepsilon))\,dG(\varepsilon) \geq \frac{\xi}{1-\beta}$;

- *well-informed* whenever $\int F(\Phi(p)) - F(\underline{\phi}(p))dG(\varepsilon)$, $\int F(\bar{\phi}(p)) - F(\Phi(p))\,dG(\varepsilon) \leq \frac{\xi}{1-\beta}$.

Using this definition, Lemma 2.4.1 can be generalized. However, Proposition 2.4.2, only holds if the taste parameter is a point-distribution as well. An analogous argument to Proposition 2.4.2 (i) can be made only if tastes are not too dispersed. Hence, the interaction between horizontal and vertical differentiation add some complications, but the forces uncovered in Section 2.4 remain present even when we allow for some limited idiosyncratic taste.

2.6 Empirical Evidence

Our analysis shows that the selection of positive versus negative WOM can depend on two factors: the strength of the brand image (how dispersed priors are) and the informativeness of negative/positive experiences. A stark and testable prediction is that with homogeneous priors, i.e. if the brand name is close to well-entrenched, no positive WOM can arise (Propositions 2.4.2 and 2.4.2). In contrast, for weak brands with "good news processes", positive WOM arises in any equilibrium (Proposition 2.4.2). In this section, we examine

these testable implications with data. The case of Yelp.com is closer to the scenario of multiple followers that we discuss in Section 2.5.3.

2.6.1 Restaurant Industry and Review Platforms

Restaurant review platforms present a good setting to validate our theoretical predictions. Restaurants are experience goods whose quality cannot be fully ascertained a priori (Nelson 1970, Luca 2016) and people often rely on recommendations from their social contacts.[18] Moreover, the restaurant industry allows us to distinguish cleanly between homogeneous and heterogeneous priors. Prior beliefs about restaurant quality naturally vary across consumers, and the extent to which consumers agree depends on how they interpret the visible characteristics of a restaurant: the brand name, cuisine, chef, etc. In this context, national chains like Subway or Domino's Pizza have invested millions of dollars to create a well-entrenched brand image with a clearly communicated brand promise and product portfolio. We can thus expect people to have homogeneous beliefs about the quality of such chain restaurants. In contrast, the industry also has smaller, independent restaurants that are typically one-store entities that cannot build such a clear reputation, and must start out with more variance in consumer beliefs about their quality. We thus expect people to have heterogeneous beliefs about the quality of independent restaurants. Existence of these two types of restaurants is critical to testing the predictions of our model. Moreover, unlike some other product categories which also have active review forums (like hotels, cars or movies), restaurants are quite local, without strong loyalty programs. Hence, it is reasonable to assume that reviewers are motivated to engage in WOM because they want to be providers of useful instrumental information, rather than by loyalty rewards or other

[18] 94 % of US diners are influenced by online reviews as per the Trip Advisor "Influences in Diner Decision-Making" survey 2018. BrightLocal's 2017 Local Consumer Review Survey estimated this number at 97 %

external incentives.[19]

2.6.2 Data Description and Summary Statistics

We construct our dataset from the Yelp Data Challenge 2017 and separate chain restaurant data. The Yelp dataset has business, review and reviewer information for restaurants in several US and some Canadian cities (majorly Pittsburgh, Charlotte, Las Vegas, Cleveland, Phoenix and Montreal) between the years 2004-2017. Every review in this dataset has a unique identifier, an overall rating, review text and timestamp. Reviews can be linked to a specific reviewer and business through unique business and reviewer identifiers. For every business, we know the name and exact location. Likewise, for every reviewer we have information like when they joined the platform, how many years they have been part of the Yelp Elite program, number of friends and fans and how many compliments they have received. We augment this dataset with other business characteristics like whether the business is a chain or not (*chain dummy*), and for chains we add the age of the brand and number of stores of the brand in US (from Statista.com and company websites). We also derive the *cuisine* variable for a restaurant using information from corporate reports for chains and name-matching for independent restaurants.[20]

The restaurants in the data cover a huge variety of cuisines; we restrict attention to cuisines for which there exist both independent restaurants and chains. We identify 72 chains and cluster them based on two dimensions, age of the chain and number of stores in the United States. Seven are classified as *national established chains* with a median brand age of *62* years and median spread of *15K* stores per chain (in US). These seven chains are Burger King, Domino's Pizza, Dunkin' Donuts, KFC, McDonald's, Pizza Hut and Sub-

[19] Yelp.com in fact recognizes this self-enhancement motive of users, and encourages users to interact and build a community, through programs like Yelp Elite. We do however find some evidence that sometimes reviewers review to give feedback to the restaurant or a particular server.

[20] Independent restaurants often have the cuisine in their name for e.g., Otaru Sushi or Mooyah Burgers. We ignore restaurants for which we cannot identify the cuisine.

way.[21] We have 30,419 reviews from 2834 such national established chain stores. There are two additional clusters that we combine in a category that we call *less established chains*. These are either old brands with limited coverage e.g., Carl's Jr and Chick-fil-A or relatively newer brands and cuisines e.g., Applebee's Neighborhood Grill & Bar, Red Lobster and Chipotle. Their median brand age is *50* years and coverage is 1000 stores across US. We have 86,359 reviews from 2913 less established chains. Most of the national established chains are sandwich, pizza, burger joints and coffee shops whereas the less established chains have a wider variety of cuisines e.g., "delis", "chinese", "breakfast" and "steak". To ensure fair comparison, we chose independent restaurants serving the same cuisines by name-matching on "sandwich", "pizza", "burger","steak", "deli", "breakfast (or brunch)", "chinese" and "coffee" categories. This gives us a total of 307,622 reviews from 6228 independent restaurants. Refer to Table 2.1 for a summary of the characteristics of the different restaurant types.

2.6.3 Supporting Evidence from Data

Rating Distribution Of Chain and Independent Restaurants

We start with describing the raw data by presenting some summary statistics, and distributions of ratings for different types of restaurants. We calculate two review statistics: the average review-level star rating and the average store-level star rating. We can see from Table 2.1 that review-level average ratings for independent restaurants tend to be higher (3.8) compared to national established chains (2.3) or less established chains (3.1). Moreover, for the independent restaurants, the average store-level star rating (3.56) is lower than average review-level star rating (3.8). Thus, "good" independent restaurants seem to receive disproportionately many reviews relative to "bad" independent restaurants. In contrast,

[21] Table 5.1 has some more details about these chains like revenue, brand value and proportion of positive, negative and neutral word of mouth

the average store-level star rating of national established chains (2.46) is higher than the average review-level star rating (2.34). Thus, "good" established chain restaurants receive disproportionately fewer reviews relative to "bad" restaurants. This difference suggests a differential propensity to review chains and independent restaurants, conditional on bad or good experiences.

We also look at the full distribution of ratings for independent restaurants and national established chains in the dataset. Figure 2.2 shows that national established chains receive a large number of 1-star reviews whereas independent restaurants receive mostly 4 and 5 star reviews. The distribution for less established chains is somewhere in between. This is consistent with our theoretical predictions: Recall that Proposition 2.4.2 suggests that in case of homogeneous priors, we should expect negative word-of-mouth. As we argued above, consumers are likely to have homogeneous prior beliefs about national established chains, and more heterogeneous beliefs for independent restaurants, and so our model would predict that national established chains have overwhelmingly negative reviews. The figure also shows the distribution of reviews separately for the first year after a restaurant appears on Yelp (light grey histogram). We do this to stay closer to our theoretical assumption of small β. The patterns are qualitatively similar.

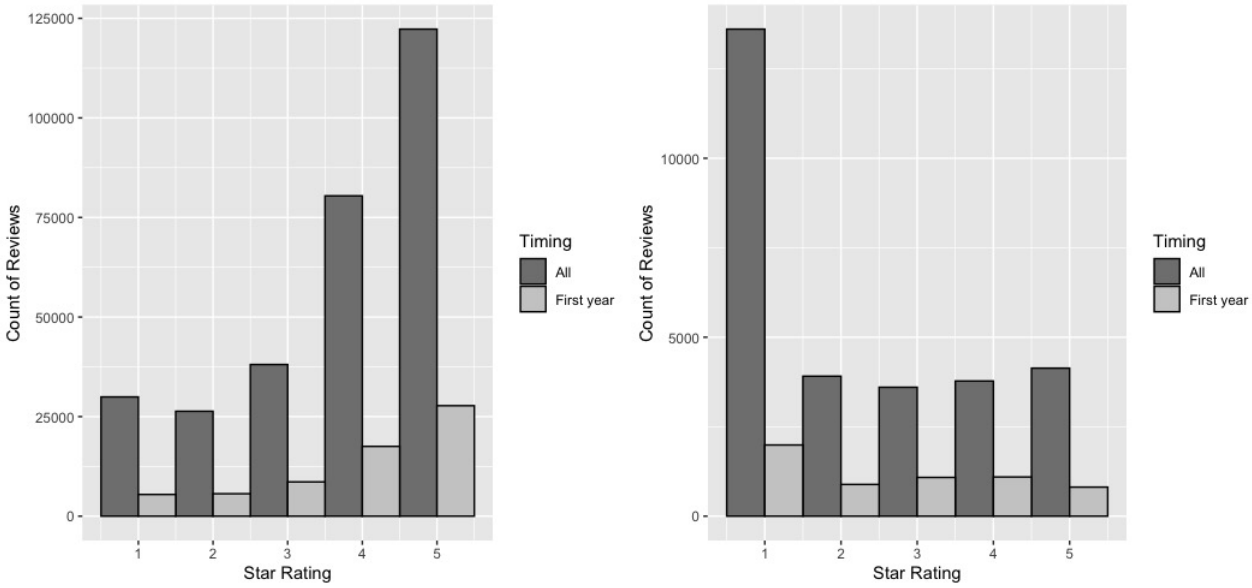
Table 2.1: Summary Statistics of Independent and Chain Restaurants

	Independent			**National Chains**			**Less Estd Chains**		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Age of Brand (Yrs)		NA		63	62	13	50	48	18
Stores in US ('000)		NA		15.8	15	5.9	1.9	1.0	2.5
Age of Store (Yrs)	3.4	3	2.9	2.9	3	2.3	3.9	4	2.8
Store Rating	3.54	3.91	0.55	2.46	2.27	0.7	2.9	3.1	0.6
Review Rating	3.8	4	1.3	2.3	2	1.5	3.1	3	1.5
No of Stores in Data		6228			2834			2913	
No of Reviews in Data		307,622			30,419			86,359	

Note: Store Rating is the average of the aggregate ratings at the individual store-level. Review rating is simply the average of all reviews. Thus, store rating gives equal weight to stores irrespective of review count. Differences in means are statistically significant ($p < 0.00001$).

A natural question is whether the difference in star ratings can be mainly attributed to

Figure 2.2: Histograms of Star Ratings: Independent versus Established National Chains



(a) Independent Restaurants (b) National Established Chains

quality differences. First, many chain restaurants repeatedly ranked higher on customer satisfaction by American Customer Satisfaction Index Survey (ACSI).[22] ACSI index and revenue data for some of the largest chains are summarized in Table 5.1 in the online appendix. Further, many of these chains have continued to show revenue and profitability growth over the years according to the Quick Service Restaurants Reports 2009-2018. Finally, the number of years that the restaurants are active in the data (in Table 2.1) is comparable across segments, suggesting that exit of low quality independent restaurants cannot explain the high average reviews of independent restaurants.

Impact of the Chain and Brand Effect on Restaurant Ratings

The differences in distributions of ratings noted above may be driven by many factors such as cuisine, location-specific heterogeneity or reviewer experience. We estimate the

[22]The American Customer Satisfaction Index (ACSI) measures the satisfaction of U.S. household consumers with the quality of products and services by surveying roughly 300,000 consumers—https://www.theacsi.org/about-acsi.

impact of being a chain restaurant on the overall rating of a restaurant, controlling for several restaurant and user characteristics — price range, cuisine, city as well as reviewer's platform experience, Elite program membership and reviewer-specific rating leniency. We specify

$$R_{ijt} = \beta_0 + \beta_1 \text{ Chain}_j + \beta_2 X_j + \beta_3 U_i + \varepsilon_{ijt} \tag{2.1}$$

where R_{ijt} denotes the rating of restaurant j by reviewer i at time t, $Chain_j$ captures whether restaurant j is a chain or not, X_j includes the restaurant price range,[23] cuisine and city, and U_i captures reviewer-specific variables such as user experience in years, an Elite dummy and reviewer average rating from other reviews. As another consistency check for our theory, we separately estimate the impact of brand age and number of stores (coverage) since these can be proxies for the strength of the brand image and can determine the dispersion of consumer beliefs.

$$R_{ijt} = \beta_0 + \beta_1 \text{ Brand age}_{jt} + \beta_3 \text{ No of stores}_{jt} + \beta_2 X_j + \beta_3 U_i + \varepsilon_{ijt} \tag{2.2}$$

Here Brand age$_{jt}$ measures the age of chain j (only applicable for chains) at time t, and no of stores$_{jt}$ is the number of stores of the chain j in US at time t. Table 2.2 (1) shows that being a chain restaurant results in getting about 1 star less than a comparable independent restaurant.[24] Further, Table 2.2 (2) shows that the propensity to write a negative review increases with age of brand and number of stores; a 50 year old brand with thousands of stores will receive 0.5 less stars as compared to a new chain with very few stores. The chain and brand age effects are quite resilient controlling for different user characteristics (4 and 5 in Table 2.2), though the magnitude of the chain effect is slightly reduced when we account for reviewer-specific leniency (average of reviewer's ratings on other

[23] Price is not the absolute price but rather a user's perception of restaurant's price range.

[24] We also ran the same regression (2.1) with only first-year reviews and the coefficients remain similar.

restaurants).[25]

We also estimate the impact of the chain effect on the number of positive, negative and neutral reviews at a business level. In particular, we run the three specifications

$$\text{Count(Rev)}_{gjt} = \beta_0 + \beta_1 \text{ Chain}_j + \beta_2 X_j + \beta_3 U_i + \varepsilon_{jt}, \tag{2.3}$$

where $Count(Rev)_{gjt}$ denotes the number of reviews of type g (positive, negative or neutral) that a restaurant j receives in the first year and over its lifetime on the platform. The reviewer characteristics U_{ij} are averaged across all reviewers of a restaurant j. The results are summarized in Table 2.3. A chain restaurant receives 8 less positive reviews in its first year than a comparable independent restaurant and 26 less positive reviews over its entire lifetime. Coupled with the fact that chains are less likely to receive any type of reviews, this is a large number of reviews and can sufficiently alter the search outcomes in a platform like Yelp.com where users rely mostly on average ratings and more recent reviews for sorting.

[

Brand Image and Beliefs: Textual Analysis of Reviews

Our premise is that the overwhelmingly negative WOM observed in chains is driven by the existence of homogeneous consumer beliefs about the brand: Negative reviews reflect deviations from what consumers collectively expect from the chain. For independent restaurants, consumers know that they do not share the same expectations, so the reference to expectations is less meaningful. If this premise is correct, a higher proportion of reviews from chain restaurants should contain words related to *expectation* or *belief* as compared to independent restaurants. Moreover, we hypothesize that these words are more likely to

[25] There could be an impact of local competition. However, it is not straightforward to define the competition set for a restaurant. So instead, we control for location(city) that captures some of this effect.

Table 2.2: Impact of Chain Dummy and Brand on Star Ratings

	Dependent variable: Overall Rating				
	rev_stars				
	(1)	(2)	(3)	(4)	(5)
Chain Dummy	-1.061***	-1.062***	-1.021***		
	(0.0641)	(0.0633)	(0.00958)		
Brand Age(Yrs)				-0.0104***	-0.0106***
				(0.000166)	(0.000159)
No of Stores (US)				-0.0000435***	-0.0000380***
				(0.000000871)	(0.000000831)
Age of Store(Yrs)	-0.0103**	-0.00834*	-0.0113***	-0.0119***	-0.0142***
	(0.00329)	(0.00365)	(0.000813)	(0.000865)	(0.000807)
Price Range					
$$	-0.144***	-0.143***	-0.130***	-0.148***	-0.133***
	(0.0338)	(0.0334)	(0.00585)	(0.00563)	(0.00543)
$$$	0.0477	0.0434	0.0725***	0.170***	0.186***
	(0.0627)	(0.0628)	(0.0128)	(0.0131)	(0.0126)
$$$	0.161*	0.153*	0.196***	0.267***	0.293***
	(0.0755)	(0.0751)	(0.0149)	(0.0151)	(0.0146)
Price Range × Chain					
$$ × Chain	0.255***	0.258***	0.255***		
	(0.0595)	(0.0588)	(0.0122)		
$$$ × Chain	0.704**	0.700**	0.654***		
	(0.245)	(0.242)	(0.0556)		
$$$ × Chain	0.724***	0.732***	0.614***		
	(0.162)	(0.156)	(0.0679)		
Select Cuisines					
burger	-0.0531	-0.0548	-0.0694**	0.184***	0.136***
	(0.0905)	(0.0895)	(0.0242)	(0.0259)	(0.0247)
chicken	-0.0563	-0.0549	-0.0232	0.0328	0.0459
	(0.0837)	(0.0831)	(0.0282)	(0.0296)	(0.0283)
chinese	-0.165*	-0.162*	-0.167***	-0.470***	-0.482***
	(0.0741)	(0.0736)	(0.0354)	(0.0378)	(0.0357)
coffee	0.296***	0.293***	0.275***	0.534***	0.482***
	(0.0732)	(0.0724)	(0.0245)	(0.0261)	(0.0249)
dessert	0.407***	0.401***	0.376***	0.314***	0.291***
	(0.0659)	(0.0654)	(0.0277)	(0.0288)	(0.0275)
pizza	-0.116	-0.114	-0.0855***	0.157***	0.143***
	(0.0750)	(0.0744)	(0.0244)	(0.0261)	(0.0249)
sandwich	0.107	0.103	0.111***	0.243***	0.218***
	(0.0668)	(0.0655)	(0.0245)	(0.0262)	(0.0251)
Reviewer characteristics					
Yelp Experience		0.0000620		-0.000126	
		(0.000229)		(0.000110)	
Elite Years		0.0263***		0.0245***	
		(0.00194)		(0.00209)	
N	418653	415423	418653	415423	418653
adj. R-sq	0.096	0.097		0.106	
User Fixed Effect	N	N	Y	N	Y

Note: Standard errors clustered by business *p<0.1; **p<0.05; ***p<0.01

Note: Restaurant controls include restaurant price range, cuisine and city, where the price range is calculated from user perceptions of a restaurant's price range. User controls include user experience in years, an Elite dummy and reviewer average rating from other reviews. To account for competition we further control for the city location of the restaurant. Specification (1) measures the chain effect without reviewer controls, (2) with reviewer controls, (3) with reviewer fixed effect, (4) and (5) measures the differential impact of brand age and no of stores for a chain brand. (4) and (5) establish that the chain effect is mainly driven by brand strength.

Table 2.3: Positive, Negative and Neutral reviews by Restaurant Type

	Dependent variable: No of Reviews by Business					
	First Year			**Lifetime**		
	Positive	Negative	Neutral	Positive	Negative	Neutral
	(1)	(2)	(3)	(4)	(5)	(6)
Chain Dummy	-7.739***	-1.033***	-1.203***	-26.46***	-0.973	-3.769***
	(0.383)	(0.133)	(0.0878)	(3.145)	(0.501)	(0.508)
Price Range						
$$	1.957***	1.149***	0.717***	5.003**	4.366***	2.480***
	(0.397)	(0.145)	(0.0832)	(1.772)	(0.397)	(0.339)
$$$	0.142	-0.0353	0.0932	23.43**	2.085	3.767**
	(0.998)	(0.336)	(0.196)	(8.330)	(1.509)	(1.170)
$$$$	2.674	0.871	0.723	85.65***	13.14**	13.97***
	(3.081)	(0.802)	(0.678)	(24.01)	(4.896)	(3.996)
Select Cuisines						
burger	2.960***	1.082***	0.814***	5.128*	-3.987***	-0.619
	(0.526)	(0.277)	(0.141)	(2.470)	(1.129)	(0.584)
chinese	-1.965**	-0.292	-0.221	-15.63***	-10.26***	-3.980***
	(0.603)	(0.379)	(0.155)	(2.092)	(1.282)	(0.575)
coffee	2.125***	-0.274	-0.207	-1.150	-7.356***	-3.332***
	(0.619)	(0.248)	(0.132)	(2.550)	(1.121)	(0.550)
mexican	-0.991*	-0.320	-0.252*	-11.16***	-10.49***	-3.974***
	(0.457)	(0.251)	(0.108)	(1.661)	(1.105)	(0.472)
pizza	-2.081***	-0.403	-0.627***	-9.794***	-6.051***	-4.226***
	(0.549)	(0.254)	(0.129)	(2.842)	(1.168)	(0.615)
sandwich	0.684	0.240	0.0956	-3.208	-8.481***	-2.749***
	(0.469)	(0.278)	(0.116)	(2.998)	(1.102)	(0.588)
Select States						
AZ	-0.980*	-0.703***	0.377***	18.58***	8.518***	3.498***
	(0.387)	(0.169)	(0.0981)	(2.205)	(0.425)	(0.354)
NC	-3.534***	-1.690***	0.0714	-3.536	-0.412	-0.408
	(0.399)	(0.162)	(0.0987)	(1.920)	(0.377)	(0.324)
NV	2.938***	0.390	1.285***	-7.280***	-1.191***	-0.141
	(0.747)	(0.332)	(0.199)	(1.951)	(0.355)	(0.320)
NY	-8.184***	-3.328***	-0.536*	-2.604	1.021*	0.398
	(0.908)	(0.369)	(0.269)	(2.201)	(0.435)	(0.373)
PA	-5.667***	-2.016***	-0.241*	1.040	2.216***	1.374**
	(0.395)	(0.169)	(0.0987)	(2.713)	(0.581)	(0.489)
Reviewer characteristics						
No of Reviews	-0.00203***	-0.000782***	-0.000135***	0.00379	-0.0000842	0.00213***
	(0.000192)	(0.0000697)	(0.0000295)	(0.00231)	(0.000473)	(0.000601)
Star Rating	2.331***	-1.053***	0.0584*	15.01***	-1.142***	1.749***
	(0.156)	(0.0571)	(0.0285)	(1.050)	(0.251)	(0.169)
N	12024	12024	12024	9719	9719	9719
R2	0.1095	0.076	0.08	0.1	0.21	0.12
adj. R-sq	0.106	0.071	0.076	0.099	0.212	0.117

Note: Standard errors clustered by business

*p<0.1; **p<0.05; ***p<0.01

be present in negative reviews of chain restaurants.

To verify these hypotheses, we examine the textual content of a subset of randomly selected 750 reviews. We are interested in how the review text differs for positive (4-5 stars), negative (1-2 stars) and neutral (3-star) reviews of chains and independent restaurants. We create a custom dictionary of *expectation* words and use it to look for instances when people mention prior beliefs and expectations in the review text. Examples of these words would be "expect", "past", "improve", "decline" to name a few. We also use the pre-built LIWC dictionary (Pennebaker 1997) to identify mentions of discrepancies in review text which capture deviation from expectations.[26] LIWC is a widely-used dictionary in psychology and marketing and examples of discrepancy words include "should", "could", "would have". Together, our custom dictionary of expectation and the LIWC discrepancy keyword list would be able to identify instances of mentions of past notions and deviations from belief. We also construct a custom dictionary of *novelty* which would identify mentions of being "surprised" or "new experiences."

Table 2.4 shows the proportion of reviews, by restaurant type and valence, that contain mentions of expectation, novelty and discrepancy. We can see that negative reviews of chains are most likely to have *expect* words (33% of all negative chain reviews). However, positive reviews of chains are also more likely to have *expect* words in comparison to independent restaurant reviews (25% versus 16-18% in independent restaurants). This is consistent with our assumption of homogeneous and strong priors for branded chain restaurants. Neutral reviews in general contain more *expect* words (which is not surprising as a 3-star most often means that the restaurant met expectations). Similarly, *discrepancy* words are more likely to be present in negative reviews. *Novel* words are most often found in positive reviews of independent restaurants and negative reviews of chains which means that people generally want to mention positive surprises for independent restaurants (for which they have uncertain priors), but report only negative surprises for chains. Interest-

[26]See Appendix Table 5.2 for our dictionaries of expectation,novelty and employee words.

ingly, employees are mentioned the most in negative chain reviews suggesting that, people often review to complain about an employee (in chains), as this is the only uncertain aspect of their visit to an established chain restaurant.

Table 2.4: Presence of Expectation, Novelty and Discrepancy Words

	Chain (Proportion of Reviews)					Independent (Proportion of Reviews)				
	N	Expect	Novel	Employee	Discrep	N	Expect	Novel	Employee	Discrep
Negative	42	0.33	0.33	0.35	0.73	90	0.18	0.21	0.25	0.75
Positive	70	0.25	0.28	0.24	0.47	378	0.16	0.33	0.2	0.61
Neutral	46	0.3	0.41	0.3	0.65	125	0.27	0.3	0.19	0.78

Expect stands for presence of expect words. Likewise, Novel, Employee and Discrep stands for presence of novelty, staff-related and discrepancy words.

Robustness: Verified Reviewers

It is widely known that fake reviews are common on many review platforms including Yelp.com (see e.g. Luca and Zervas (2016)). While the long-term negative impact of fake reviews seems to be limited (see He et al. (2020)), Mayzlin et al. (2014) document that, in the hotel industry, chains have a higher propensity of receiving fake negative reviews when the neighborhood includes more independent hotels.

Our dataset excludes all reviews that were identified by the Yelp filter to be fake, but the filter is likely not able to filter out all fake reviews. To show that the observed differences in valence cannot be completely attributed to chains receiving more fake negative reviews, we re-run our analysis with a subset of reviews written by Yelp-verified Elite Reviewers, which are guaranteed to be genuine. First note that Table 2.5 shows that there are no major differences between Elites and Non-Elites in the type of restaurants they reviewed: Both groups review an almost equal proportion of chains and high-end restaurants (at least for the cuisines we are studying i.e. “sandwich”, “pizza”, “burger”, “delis”, “coffee” etc. mentioned earlier). Elites, do write slightly more positive reviews (average Elite rating is 3.7 compared to 3.4 for Non Elites) and tend to review newer restaurants (average age of

restaurant reviewed by Elites is 3.1 years whereas for Non Elites it is 3.6 years). Table 2.6 summarizes the analogous results to Table 2.2 with only Elite reviews. In spite of the fact that Elites write more positive reviews, the reviews for chain restaurants are still 0.5 stars lower than comparable independent restaurants. Thus, our results are robust and remain qualitatively similar for a pool of verified reviewers.

Table 2.5: Elite and Non Elite Reviewers

	Rating (Mean)	StoreAge (Mean)	Experience (Mean)	% Chains	% High-End
Elite	3.7	3.1	82	26.3%	10.7%
Non Elite	3.4	3.6	61	27.1%	10.4%

2.7 Conclusion

We propose a theoretical model of strategic WOM that explains how positive and negative WOM arises in equilibrium. We highlight two factors that determine selection of positive versus negative WOM — the strength of the brand image as measured by the dispersion of beliefs about quality, and the informativeness of good and bad experiences. The brand image affects how many customers the firm can attract given its profit-maximizing price, which in-turn impacts how many consumers can be influenced by WOM.

On platforms like Yelp.com, users rely mostly on average ratings to sort. A practical implication of our results is that since the propensity to review varies after good or bad experiences based on the brand image, average reviews are not a reliable measure of quality [27]. More specifically, WOM needs to be interpreted differently for different types of restaurants, and it can be problematic to use only rating comparisons on review platforms to make purchasing decisions. Solutions can be to incentivize all consumers to write

[27] Jin et al. (2018) also highlight the disadvantages of focusing on average ratings alone and define an adjusted average that accounts for reviewer heterogeneity and past ratings

Table 2.6: Impact of Chain Dummy and Brand on Star Ratings (For Elites)

	Dependent variable:			
	rev_stars			
	OLS			
	(1)	(2)	(3)	(4)
Chain Dummy	-0.539***	-0.538***		
	(0.012)	(0.012)		
Brand Age (Yrs)			-0.001**	-0.001**
			(0.0005)	(0.0005)
No of Stores			-0.00003***	-0.00003***
			(0.00000)	(0.00000)
Age of Store (Yrs)	0.011***	0.005***	0.008***	0.002
	(0.001)	(0.001)	(0.001)	(0.001)
Price Range				
$$	-0.107***	-0.108***	-0.106***	-0.107***
	(0.010)	(0.010)	(0.010)	(0.010)
$$$	0.147***	0.158***	0.185***	0.198***
	(0.021)	(0.021)	(0.021)	(0.021)
$$$$	0.334***	0.356***	0.373***	0.396***
	(0.024)	(0.024)	(0.024)	(0.024)
Observations	90,588	90,588	90,588	90,588
R^2	0.064	0.069	0.073	0.077
User Characteristics	N	Y	N	Y
Note:			*p<0.1; **p<0.05; ***p<0.01	

reviews, or to present more sophisticated aggregated ratings that control for systematic selection in reviews.

Finally, our research has important implications for understanding the link between "conversational motives" and outcomes like valence. We find that the text in the reviews can help identify the motivation of the reviewer (expectation deviance or reporting novel experiences). Text analysis can be useful more generally to identify drivers of selection issues in reviews, and to control for them.

We leave the questions around optimal design of review aggregation mechanisms and a broader understanding of WOM motives for future research.

Chapter 3

What makes sales pitches work? Using Multi-Modal Video Data to Uncover Success Factors

3.1 Introduction

Sales force recruitment and training is a key spending area for US firms across industries. A Deloitte 2016 study finds that the cost-per-hire metric for US firms has gone as high as $4000 [1]. Moreover, in complex jobs like sales, there is also a significant ramp-up cost for new employee on-boarding. US firms spend close to 15 billion each year on salesforce training and reducing salesperson onboarding time remains one of the major concerns of Chief Sales Officers across the country [2]. This is especially critical given the high attrition rate for salespeople across organizations which means frequent hiring and training expenditures[3]. Inspite of proliferation of new types of selling channels, the sales pitch which comprises of a buyer-seller limited-time interaction still remains an important

[1] https://www.prnewswire.com/news-releases/bersin-by-deloitte-us-spending-on-recruitment-rises-driven-by-increased-competition-for-critical-talent-300070986.html

[2] https://www.brainshark.com/sites/default/files/cso-insights-2016-sales-enablement-optimization-study.pdf

[3] The 2016 Sales Performance Optimization Study finds that the average turnover rate is 16.3 % and more than 40 % of organizations report that they need more than 10 months to ramp up a new sales personnel

element of the selling process. Often times, the sales pitch is the first interaction of a buyer with a company and the salesperson's behavior determines the long-term financial relationship between two firms. Hence, companies spend a lot of resources to train their salesforce in mastering sales pitches using a mix of classroom, field and role-play oriented training programs. Moreover, a key component of recruitment or training programs is evaluation i.e., companies need in-house or external experts to recruit the right candidates and evaluate their progress through the training process. In this paper, we propose to develop an automatic salesforce assessment tool combining the rich literature on sales tactics and persuasion in marketing and the recent advances in the areas of multi-modal data analysis (textual, audio and visual) in computer science. Our research objectives are two fold: First, we want to develop an automated assessment tool that can be used for both initial screening of salespersons during recruitment as well as monitoring their progress throughout the sales training program. Second, we want to advance the personal selling literature by identifying underlying seller behaviors that lead to the success of certain influence tactics with certain types of buyers using videos that capture the entire sales interaction.

3.2 Related Literature

Our work is connected to several strands of literature — the marketing and personal selling literature on influence tactics in sales, the psychology and economics literature on persuasion and the current computer science literature that connects audio and textual elements to personality and orientation.

Our salesforce assessment tool uses a feature set of variables derived from the marketing literature on personal selling (Sheth 1976, Frazier and Summers 1984, Spiro and Weitz 1990) that has identified the importance of content, style and adaptability (similarity between buyer and seller) as the most important factors determining sales success. Content

is defined as the actual product utilities that the seller offers and buyer demands while style is defined as the "format, ritual and mannerism" that the buyer and seller adopt in their interaction. The definition of style has been more abstract and there are no unifying models of style, however, it is largely understood as elements of language (e.g., noun-verb ratio, no of personal pronouns used) as well as some non-lingustic elements like speech rate, energy in the voice as well as kinesics like hand movements and gestures (Williams and Spiro 1985, Pennebaker et al. 2001). These variables are extremely hard to measure in the absence of recorded videos of buyer-seller interactions and technology to precisely measure these constructs. Likewise, research on similarity or adaptability has been largely inconclusive (Churchill et al. 1975, Evans 1963) due to the absence of similarity measures beyond the demographic characteristics of buyers and sellers.

The main contribution of our paper lies in quantifying the impact of communication style (of the buyer-seller interaction) and buyer-seller similarity using textual and audio data while controlling for content and characteristics of the dyad. Also, we will study the impact of style at different stages of a sales pitch. Wilson (1976) is one of the early papers that lays out the different stages of a buyer-seller interaction namely source legitimization, information exchange or problem identification, attribute delineation, attribute value negotiation and relationship maintenance. In line with this, we measure impact of content, style and similarity during introduction, need identification, presentation, objection handling and closure phases of a buyer-seller negotiation. Different tactics could be more important at different stages of the sales pitch. For example, being confident and energetic at the start could be effective but probably a more calm, interactive approach works during the display stage and the later objection handling phases.

Non-verbal communication has been extensively studied in behavioral sciences (Jones and LeBaron 2002, Mehrabian 2017). Some of these studies have also studied the relative importance of different modalities like voice or text for persuasion outcomes (Scherer et al. 1973, Van Zant and Berger 2020, Wang et al. 2021). Our work also compliments some

recent studies on persuasion in computer science (Longpre et al. 2019, Shmueli-Scheuer et al. 2019) and business (Manzoor et al. 2020) that look at textual and audio-visual data. However, these studies have typically looked at persuasion as a one-shot construct and have not focused on different stages of the process. Also, most of these studies focussed on persuasion in the domain of social media videos that are typically one-sided. On the other hand, a buyer-seller interaction is a two-way communication where the role of interactivity and adaptability is a key factor over and above the message content and delivery (which are both characteristics of the persuader).

3.3 Background and Data

In order to build a predictive model of persuasion for buyer-seller interactions, we need training data that has been pre-labeled by human taggers. We use a unique data source of buyer-seller role-plays that give us the flexibility of a lab study while giving incentives to both buyers and sellers to put their best foot forward. Our data comes from the sales lab of a large research institution where student sellers engage in a 20 min conversation with a buyer who is a seasoned sales professional within their organization. During this conversation, the seller tries to persuade the buyer to sign a contract with their company for a technology related product. The duration of the interaction is fixed and sellers and buyers are both briefed about the product and scenario beforehand. This interaction is not only a grading component for the students but can also result in an internship with the buyer's firm. Thus, the buyers take part in this role-play to evaluate good prospective salespeople. Thus, both buyers and sellers have a real incentive to participate in this interaction.

During the exchange, the buyer-seller interactions are rated by experienced judges (who are sales professors at the university) on the National Collegiate Sales judgement criteria (See Fig 3.1). The judges rate the sellers based on the following stages of the interaction:

1. Introduction: This is the stage where the seller greets the buyer and tries to engage in small talk in order to understand the buyer better and build a rapport. This stage typically comprises of the first 2-3 mins of a conversation

2. Need Identification: This stage follows the introduction stage. Here, the seller digs deeper into the problem areas for the buyer's organization and tries to uncover implicit and explicit needs. Sellers typically spend 6-8 minutes for need identification

3. Presentation: This is where sellers present their product to the buyers, usually they would give a demo of the software and describe key features and how it can meet the needs of their organization. This stage typically consists of 2-5 minutes.

4. Objection Handling: In this stage, the buyer raises objections about the product and the seller has to put forth counter arguments. The length of this stage is normally 4-5 minutes, however sometimes this continues for longer and the seller runs out of time.

5. Closure: In this stage, the buyer makes a decision and the duo decides on next steps

The judges also give 3 overall scores : persuasion (whether the seller makes a persuasive pitch), confidence (overall confidence level of the seller) and enthusiasm (did the seller come across as energetic). All the scores are on a scale of 1-10 where 10 is the highest score. Our data comprises of all the judge scores as well as the actual videos of the buyer-seller interaction. We then use AWS Transcribe to extract time-stamped text transcripts of this exchange. Thus, the full dataset is multi-modal — it includes video (which has both audio and visual components) as well as text (from the transcripts) and an associated judge scoring matrix.

3.3.1 Buyer Types

We also tag the videos to capture some buyer characteristics like gender, experience and orientation. We use the definitions of buyer orientations as described in Sheth (1976). A relationship oriented buyer is warm, friendly and cares more about building a long-term relationship through the sales process. On the other hand, a transaction oriented buyer is more task-focused and goal-oriented and likely to have a more serious demeanour. We use human taggers to evaluate the interactions and categorize buyers into relationship and task-oriented.

3.3.2 Descriptive statistics

We use 190 videos that comprise of the finale interviews across the years 2018-19 (the toughest part of the college challenge). This is done to ensure that the interactions are very close to real sales conversations. We have a unique seller in each video and 24 buyers who take turns to interact with the students. These videos are then evaluated by a panel of expert judges (each video is evaluated by 9-10 judges). Overall, 261 expert judges have scored these videos resulting in 1752 evaluations. The inter-judge agreeability for the persuasion score is high (Krippendorff's alpha = 0.72) and thus we can use this as a reliable outcome variable to train an ML model. Table 3.1 shows some summary statistics of the scoring dataset. The judges score the videos on a scale of 1- 10 (however, the minimum score is 7 across all our videos). For the purpose of building a binary predictive model of persuasion, we segment videos into *high persuasive* and *low persuasive videos*. High Persuasive videos have a score greater than the median score (8.5). These are also the students who usually land the internships. Like we can see in Table 3.1, we have 94 high persuasion videos and 97 low persuasion videos in our dataset which ensures a good class balance. Overall, high persuasion pitches have higher scores across all dimensions and especially for the objection handling and display stages. We now look at differences

Figure 3.1: NSC Judgement Criteria

National Collegiate NCSC Sales Competition

National Collegiate Sales Competition
Contestant Evaluation Form

Instructions
- Use a #2 pencil only
- Make dark marks.
- Fill in a response like this: ⓪ ● ②
- Erase completely to change, then remark.

SESSION NUMBER	
⓪①②③④⑤⑥⑦⑧⑨	
⓪①②③④⑤⑥⑦⑧⑨	

COMPETITOR NUMBER	
⓪①②③④⑤⑥⑦⑧⑨	
⓪①②③④⑤⑥⑦⑧⑨	
⓪①②③④⑤⑥⑦⑧⑨	
⓪①②③④⑤⑥⑦⑧⑨	
⓪①②③④⑤⑥⑦⑧⑨	

ADJUDICATOR NUMBER	
⓪①②③④⑤⑥⑦⑧⑨	
⓪①②③④⑤⑥⑦⑧⑨	
⓪①②③④⑤⑥⑦⑧⑨	

COMPETITOR NAME

Approach (Effectively gains attention and builds rapport)	5%	*Approach*
Professional introduction	⓪①②③④⑤⑥⑦⑧	
Salesperson gains prospect's attention	⓪①②③④⑤⑥⑦⑧	
Effectively builds rapport	⓪①②③④⑤⑥⑦⑧	
Smooth transition into needs identification	⓪①②③④⑤⑥⑦⑧	
Needs Identification (Obtains clear understanding of customer's situation)	25%	*Needs Identification*
Uncovered decision process (decision criteria, people involved in decision process)	⓪①②③④⑤⑥⑦⑧	
Effectively determined relevant facts about company and/or buyer	⓪①②③④⑤⑥⑦⑧	
Effectively uncovered needs of buyer (discovered current problems, goals, etc.)	⓪①②③④⑤⑥⑦⑧	
Asked effective questions that brought to buyers' attention what happens to company or buyer when problems continue (helped convert implied to explicit needs)	⓪①②③④⑤⑥⑦⑧	
Gain pre-commitment to consider the product/service and smooth transition to presention	⓪①②③④⑤⑥⑦⑧	
Product/Service Presentation (Persuasively match product benefits to meet buyer needs)	25%	*Product/Service Presentation*
Presented benefits-based upon needs of buyer instead of only features	⓪①②③④⑤⑥⑦⑧	
Logical, convincing presentation (display a strategy to communicate and persuade; clearly understand needs "hot buttons" of prospect and concentrates on those needs)	⓪①②③④⑤⑥⑦⑧	
Used appropriate/professional visual aids	⓪①②③④⑤⑥⑦⑧	
Effectively demonstrated product/service	⓪①②③④⑤⑥⑦⑧	
Effectively involves the buyer in the demonstration	⓪①②③④⑤⑥⑦⑧	
Effective use of trial closes (follow-up to determine where buyer is in decision process)	⓪①②③④⑤⑥⑦⑧	
Overcoming Objections (Eliminate concerns to the customer's satisfaction)	15%	*Overcoming Objections*
Initially gains better understanding of objection (clarifies or allows buyer to clarify objection)	⓪①②③④⑤⑥⑦⑧	
Effectively answers the objection	⓪①②③④⑤⑥⑦⑧	
Confirms that the objection is no longer a concern of the buyer	⓪①②③④⑤⑥⑦⑧	
Close (Take initiative to understand where you stand with buyer now and in the future)	10%	*Close*
Persuasive in presenting a reason to buy	⓪①②③④⑤⑥⑦⑧	
Asked for business or appropriate commitment from buyer, given the nature of this particular sales call	⓪①②③④⑤⑥⑦⑧	
Communication Skills	15%	*Communication Skills*
Effective verbal communication skills (active listening; restated, rephrased, clarified, probed for better understanding)	⓪①②③④⑤⑥⑦⑧	
Appropriate non-verbal communication	⓪①②③④⑤⑥⑦⑧	
Verbiage (clear, concise, professional)	⓪①②③④⑤⑥⑦⑧	
Overall (Take initiative to understand where you stand with buyer now and in the future)	5%	*Overall*
Salesperson enthusiasm and confidence	⓪①②③④⑤⑥⑦⑧	
Product knowledge	⓪①②③④⑤⑥⑦⑧	

3735
3/17

in interactivity and some select linguistic and audio features in high and low persuasive pitches in Table 3.2. We discuss these features in more detail in the empirical framework section. Successful pitches are in general more interactive (higher average number of turns). They involve fewer long monologues by the buyer or seller. The word count is somewhat higher for the successful pitches and they are more energetic.

Table 3.1: Persuasive versus Non-Persuasive Videos (Content)

	Mean (SD)						
Type of Video	Intro	Needs	Display	Objections	Communication	Confidence	N
Persuasion (High)	8.9 (0.38)	8.9 (0.35)	9 (0.42)	8.7 (0.44)	9.1 (0.32)	9.25 (0.39)	94.00
Persuasion (Low)	8.5 (0.43)	8.5 (0.44)	8.4 (0.52)	8.0 (0.43)	8.6 (0.42)	8.57 (0.54)	97.00
No of sellers	337						
No of buyers	24						
No of judges	381						

Table 3.2: Interactivity and Textual/Audio Features

Type of Video	Turns	Buyer_M (sec)	Seller_M (sec)	Buyer_WC	Seller_WC	Energy
Persuasion (High)	58 (14)	76 (34)	46 (17)	1349 (321)	1432 (313)	0.05 (0.03)
Persuasion (Low)	54 (12)	89 (43)	53 (89)	1270 (319)	1373 (322)	0.04 (0.02)

Buyer(Seller) *M* is the maximum duration of a buyer (seller) monologue. Buyer(Seller) *WC* is the total number of words spoken by the buyer and seller respectively. These measures are derived from the transcript of the conversations. Energy is captured from the voice of the seller.

3.4 Empirical Framework

The prediction outcome variable of interest is the binary persuasive construct (High or low persuasion). The gold standard is a model that predicts whether an interaction is persuasive or not based on the scores of the judges. Our goal is to build a predictive model that incorporates the textual, audio and visual features in our video data to come up with a persuasion scoring tool that matches the accuracy of human raters.

We hypothesize that four types of features impact persuasion: content of the message, style of delivery, interactivity and similarity between buyer and seller (in terms of static

and stylistic elements). The first two feature groups (content and style) are seller-driven whereas the third and fourth feature groups (interactivity and similarity) are a function of both buyer as well as seller.

- Content: This comprises of the substantive, topic-oriented aspects of the message excluding any stylistic elements. We run a topic model on our entire conversation transcript data and find 4 important topics that are discussed in these interactions. They are closely related to the four different stages of the conversation. These include topics around *Greeting and Pleasantries*. These generally occur during the beginning of the conversation and involve words and phrases like "Hey", "Good Morning", "weather". As the conversation proceeds, the other topics that get mentioned can be categorized into *Busines*, *Technology* and *Pricing*. The *Business* topic is related to the buyer's business problem and include words like "salesperson", "manager", "growth" and "profitability". The *Technology* topic is about the more technical aspects of a product (as the scenario involves selling a technology product) and consists of words like "app","cloud", "infrastructure" to name a few. *Pricing* topic is about the cost of set-up and payment terms and conditions. In Table 3.3 we summarize the top words for each of the topics. The content-related features include the proportion of each topic mentioned by a seller during the interaction.

- Style: This comprises of both verbal as well as non-verbal communication cues. Hence, it is extracted from textual, audio and visual features of the data. Textual features that are indicative of style include verbosity (no of words per sentence), non-verb ratio, various parts of speech used (e,g., prepositions, verbs, adverbs), positive and negative emotion words, words signifying certainty as well as tentativeness (e.g., absolutely, perfectly, obviously, maybe, could be, try), assent words (yeah, yes, of course) to name a few. We use the LIWC dictionary (Pennebaker et al. 2001) to extract these stylistic elements from the transcripts of the interaction. LIWC has

Table 3.3: Most common words for each content class

Technology	Business	Greeting	Pricing
laptop	salesforce	come	chargeable
security	decision makers	hey	buy
mobile	business	morning	prices
pipeline	leads	meet	discount
data	quoting	weather	monetary
storage	customers	sunny	expenditure
cloud	sales	week	subscriptions
digital	communication	thank you	budgeted
transformation	employees	yankee	wallet
electronic	problems	holidays	cost
screenshot	territories	spring break	expensive
licenses	ownership	christmas	pay
firewalls	people	business card	dollars
emails	agents	traditions	worth
video	company	greeted	price
licensing	potential client	thanksgiving	money
website	insurance	cooking	
touchscreen	manager	enjoyed	
login	oversee	breakfast	
leaderboards	monthly reports	Friday	
printout	team	family	
technologies	region		
usb	visibility		
kiosk	growth		
android	profitability		
versions	promotion		
charts	revenues		
database	forecasting		
mobility	goal		
phone	supervisors		
app	premium		
tablet	competitive		
dashboard	due diligence		
computer	reps		
tool	trajectory		
	elaborating		
	underperforming		
	policy holder		

been widely used in social sciences to quantify stylistic elements of language. In voice, style is mainly conveyed by energy, pitch, speech rate and voice modulation. We use the pyaudio library in Python to extract voice features. Visual elements that can communicate style include hand movements (velocity and amplitude) as well as body postures.

- Interactivity: This measures whether the dialogue involves active participation from both seller and buyer. A highly interactive conversation would include higher no of turns and would not have long monologues from either the buyer or seller.

    1. Turns: This is the simply the no of times the conversation moves from buyer to seller and vice-versa.
    2. Buyer Seller Share of Voice (SOV): This is the ratio of the speaking time of the buyer versus the speaking time of the seller. Even if the conversation has many turns, it might be the case that one speaker has long monologues whereas the other speaker just nods or says shorter sentences. Thus, over and above turns, this metric captures the participation of both speakers.
    3. Buyer (Seller) Max Monologue: This captures the longest time duration during which either party speaks uninterrupted.

- Similarity : We measure similarity in terms of both static characteristics of buyer and seller (e.g., gender and age) as well as stylistic factors that we notice during the interaction. The style similarity between buyer and seller is operationalized as the Euclidean distance between the style vectors of the buyer and the seller where each style vector has language and audio-visual elements extracted from the interaction videos.

Table 3.4 summarizes the textual, audio and interactivity features that we extract for quantifying linguistic style and measuring stylistic similarity between the buyer and the seller.

Table 3.4: Textual, Audio and Interactivity Features

Type	Modality	Feature	Explanation
Style	Text	Words per sentence	This measures how verbose the sentences are
Style	Text	Complex words	Using words that have more than 6 letters
Style	Text	Parts of speech	Proportion of verbs, aux verbs, adjectives and prepositions
Style	Text	Individualistic	“I”, “I’m”, “me”, “my”
Style	Text	Collaborative	“we”, “us”, ”both”
Style	Text	Assent	“yes”, “obviously”, “yeah”
Style	Text	Filler Words	“hmm”, “aah”,“huh”
Style	Text	Certainty	“ofcourse”, ”definitely”, ”absolutely”,”confident”, “sure”
Style	Text	Tentativeness	“maybe”, ” could be”, ” likely”
Style	Text	Emotional Words	Positive and Negative Sentiments
Style	Text	Politeness	“please”, “Thanks”, “grateful”
Style	Text	Questions	Question Marks, “how”,“when”, “Where”
Style	Text	Thinking/Insight	“think”, “most likely”
Style	Text	Cause	“because”, “reason”, “causes”, “therefore”
Style	Text	Difference	“but”, “yet”,” disagree”
Style	Text	Achieve	“success”, “win”, “launch”, “therefore”
Style	Audio	Energy	Measures how energetic or enthusiastic the seller is
Style	Audio	Speech Rate	No of words per sec
Style	Audio	Brightness	The Spectral Centroid measures whether the voice is bright or dull
Style	Audio	Voice Modulation	No of times the voice signal goes from low to high and vice versa
Interactivity	NA	Turns	No of turns in the buyer-seller conversation
Interactivity	Audio	Buyer_Max_Monologue	The maximum duration (in secs) during which the buyer talks uninterrupted
Interactivity	Audio	Seller_Max_Monologue	The maximum duration (in secs) during which the seller talks uninterrupted
Interactivity	Audio	buyer_Seller_SOV	Ratio of time buyer speaks to seller speaks

3.5 Results

In this section, we discuss the performance of models that use different combinations of features for prediction. We try a range of predictive machine learning algorithms like SVM, Logistic Regression and Random Forest but report the Logistic Regression results as it is the best performing model. We then do a variable importance study to understand what features are driving the outcome decisions of the machine learning model. Finally, to understand the heterogeneous impact of different content, style and interactivity related measures on different types of buyers, we repeat the analysis for specific groups of buyers (Male vs Female buyers and Relationship-oriented vs Transaction Oriented buyers).

In Table 3.5, we summarize the performance of the Logistic Regression based models. The gold standard model is the one that predicts persuasiveness based on the scores that judges give at different stages of the interaction. Such a model is able to predict persuasion with an 82 % accuracy. Since it is a binary classification task, the baseline accuracy is 50% which is the probability of identifying the correct class by random guessing. We fist build simple models that capture only one set of features (content, style, interactivity). The model that only uses content features is as good as a random guessing model. However, models that incorporate stylistic and interactive features perform better on accuracy, precision as well as recall measures. The model based on stylistic features achieves an F1 score of 61%. The audio features improve the prediction accuracy of the model and combining all textual and audio features gives an accuracy of 63 % (i.e. 26 % improvement over random guessing).

Table 3.5: Performance Comparison

Model	Feature Set	Accuracy	Precision	Recall	F1
Logistic Regression	Content	50%	50%	52%	51%
Logistic Regression	Style	61%	59%	70%	64%
Logistic Regression	Interactivity	56%	54%	85%	66%
Logistic Regression	Style + Content + Interactivity (Only Text)	54%	52%	70%	60%
Logistic Regression	Audio	57%	55%	57%	56%
Logistic Regression	Style + Content + Interactivity (Text+ Audio)	60%	59%	63%	61%
Logistic Regression	All Features (Text + Audio)	63%	60%	74%	68%
Logistic Regression	Judge Scores (Gold standard)	83%	82%	82%	82%

In Table 3.6, we describe the most important variables by models. In the prediction model based on scores of human judges, the objection handling and display stages are mainly driving the prediction outcomes which means these are the most important stages of the buyer-seller interaction. Among content-based features, topics related to pricing and business are most important. Among stylistic features in the text, *Buyer certainty* (that captures buyer's confidence and authoritativeness) is the most predictive feature followed by the use of adjectives, collaborative words (like "we") and words that signal agreement ("yeah","yes"). Among the audio features, the most predictive feature is the Energy in the first 5 minutes of the conversation followed by the overall energy during the interaction.

In the model that takes all textual, audio and interactive elements as input, the most predictive features are conversation turns, overall Energy, content around Pricing and Buyer's certainty (which could be interpreted as confidence).

Table 3.6: Variable Importance

Model	Most Important Features (Variable Importance Measure)
Judge Scores	Objection(0.87), Display (0.62), Non Verbal (0.31), Verbal Comm (0.28), Introduction (0.23)
Content	Price (0.17), Business (0.11)
Style	Buyer_certain (0.57), Buyer_adj (0.56),Buyer_we (0.31), Buyer_assent (0.11)
Audio	Energy_5 (0.144), Energy (0.002)
All Features	Turns (Interactivity), Energy (Voice), Price (Content), Buyer_Certainty(Style)

We now split the data into interactions that involve male buyers and female buyers. We then run the same model on these two subsets of data and derive variable importance measures. Likewise, we run the same model on two subsets of data that include transaction-oriented versus relationship oriented buyers. This analysis helps us to understand if factors that lead to persuasiveness are different for different types of buyers. Table 3.8 summarizes the key insights. We find that male buyers are more content driven; especially pricing related discussions are very important to persuade them. They value a style that is both confident (seller shows certainty) as well interactive. It is important for the seller to be energetic in the first 5 minutes of the conversation. For female buyers, the more important features are related to style and interactivity. They especially like a collaborate tone which is energetic and bright.

Relationship oriented buyers care more about interactivity. While both types value style, they prefer different types of styles. Relationship oriented buyers like a warm and interactive style. The energy at the start of the conversation is the most important feature for them. Overall, they value interactivity a lot and prefer the use of "we" words and like to hear questions. On the other hand, the transaction oriented buyers like an insightful, rational and confident style shich involves questions and comparisons. They also like to

hear achievement related words like “win”, “success” and “goal”.

Table 3.7: Variable Importance by Buyer Type (I)

Male Buyers	Female Buyers
Pricing (Content)	Cause (Text Style)
Certainty (Text Style)	Energy (Audio Style)
Energy in first 5 mins (Audio Style)	“We” words (Text Style)
Assent/Agreement (Text Style)	Certainty (Text Style)
Buyer Seller SOV (Interactivity)	Voice Brightness (Audio Style)
Technology (Content)	Drive (Text Style)
Insight (Text Style)	Greeting (Content)
Business (Content)	Questions (Interactivity)

Table 3.8: Variable Importance by Buyer Type (II)

Relationship Oriented	Transaction Oriented
Energy in first 5 mins (Audio Style)	Insight (Text Style)
Certainty (Text Style)	“You” and “I” words (Text Style)
Energy in first 5 mins (Audio Style)	“We” words (Text Style)
Assent/Agreement (Text Style)	Filler words (Text Style)
Technology (Content)	Cause (Text Style)
Questions (Interactivity)	Comparison (Text Style)
Buyer Seller SOV (Interactivity)	Questions (Interactivity)
“We” words (Text Style)	Buyer Seller SOV (Interactivity)
Cause (Text Style)	Achieve words (Text Style)

3.6 Conclusion and Future Work

Overall, we find that combining different modalities (like text, audio, interactivity) improves the prediction of persuasiveness for buyer seller transactions. Using our prediction models and variable importance measures, we can see that being interactive, confident (certain), collaborative and energetic are very important for the success of a sales pitch while there are some interesting differences based on buyer gender and orientation. We think that incorporating visual features (like hand movements and posture) would improve the prediction power of our models. We also believe that coming up with a metric of sim-

ilarity (for buyer and seller style) and using this as a feature in our model can improve performance. Finally, we want to do this analysis at every stage of the negotiation process to specifically understand what types of behaviors lead to success at different stages. This would be an important incremental contribution over existing work in this domain. We wish to incorporate these ideas into our future work.

Chapter 4

Appendix to Chapter 1

4.1 Mturk Experiment

In this section, we describe the Mturk Experiment we run to motivate the importance of attribute level ratings. First, to establish that enhanced ratings are useful for customers to make better decisions, we conducted a 2×2 between subjects study on MTurk with 165 participants. Both the treatment and control groups are shown 4 restaurant reviews and asked to chose a restaurant. Every restaurant is extremely good at one of the attributes— food, service, price or ambiance and average on other attributes. The only additional information given to the treatment group is enhanced attribute level ratings. See Fig 4.1 for details of the study design. We compare the treatment and control groups on two parameters— match and attention. We consider a match when a person's restaurant choice matches with their separately elicited preference i.e. a person who says she values food chooses the restaurant that has excellent food and so on. We get our measure of attention based on whether the survey respondent correctly answers the attention check question (Q: How many restaurant choices did you have in the previous question ?) that we ask soon after the restaurant choice question. We show in Table 4.1 that providing attribute sentiment scores in addition to text significantly improves the ability of customers to choose

restaurants consistent with their separately elicited preferences over restaurant attributes. There is also a significant positive impact on attention. The fact that the treatment group is more attentive and makes choices more consistent with preferences shows that attribute level ratings reduce the cognitive burden of consumers and help them in decision making.

Table 4.1: Match and Attention Comparison: Treatment (Attribute Scores)

	N	Mean (SD)	
		Match	Attention
Treatment	74	0.7 (0.46)	0.94 (0.46)
Control	90	0.38 (0.49)	0.83(0.49)
		$p<0.01$	$p<0.05$

4.1.1 LDA

We use document level LDA as an exploratory study of which topics are discussed in reviews. We also use SLDA (supervised LDA) for extracting better topics. LDA and SLDA results are shown below in Fig 4.2

4.2 A Hybrid CNN-LSTM Deep Learning Architecture

In this section, we include a more detailed discussion of the two most important layers of the hybrid CNN-LSTM: the convolutional layer and the long short term memory layer.

Convolution Layer. The first feature generating layer in our architecture that follows the embedding layer is the convolution layer. Convolution refers to a cross-correlation operation that captures the interactions between a variable sized input and a fixed size weight matrix called filter (Goodfellow et al. 2016). A convolutional layer is a collection of several filters where each filter is a weight matrix that extracts a particular feature of the data. In the context of text classification, a filter could be extracting features like bi-grams that stand for negation *e.g. not good* or unigrams that stand for a particular attribute e.g. chicken. The two key ideas in a convolutional neural network are weight-

Figure 4.1: Mturk experiment: Importance of attribute sentiments

Q4

Imagine you are planning your next dinner to an Italian fine dining restaurant. Italian restaurants typically have a varied menu consisting of specialty pizzas, pastas and range of chicken, fish and other meat entrees.

Below are reviews of 4 different Italian restaurants in your city (each review is for a different restaurant). Read the reviews carefully and chose the restaurant that you would most likely go to!

Rating : 4/5
Food: 5 Ambiance: 3 Price: 3 Location: 2 Service: 3
Great place for an Italian dinner. We started with some house pasta and ceaser salad which were outstanding. One of the best pastas I have had. They have an outstanding dessert menu- I would not forget the chocolate cake with gelato ice-cream!The atmosphere is classy but relaxed. The service is quite decent given the amount of people they have every day. Parking can be painful but it was a Friday night and the whole city is out partying, should have Ubered!! I found the prices reasonable though the desserts were a bit expensive

Rating : 4/5
Ambiance: 5 Food: 3 Service: 3 Price: 3 Location: 2
The look and feel of this Italian diner is amazing! Wow!! Such beautiful décor and such a large and comfortable seating area. The tables are large and the dinner area is quite airy. The food at most is ordinary, though certainly not bad - I am a big fan of Italian food and have had better pastas and ceasar salad. The prices are on the higher side but cmon this is a special place! There is a parking lot but it fills up too fast and Saturday nights can be a parking nightmare! The servers are nice people, kind and attentive but nothing extra ordinary

(a) Treatment Group

Q5

Imagine you are planning your next dinner to an Italian fine dining restaurant. Italian restaurants typically have a varied menu consisting of specialty pizzas, pastas and range of chicken, fish and other meat entrees.

Below are reviews of 4 different Italian restaurants in your city (each review is for a different restaurant). Read the reviews carefully and chose the restaurant that you would most likely go to!

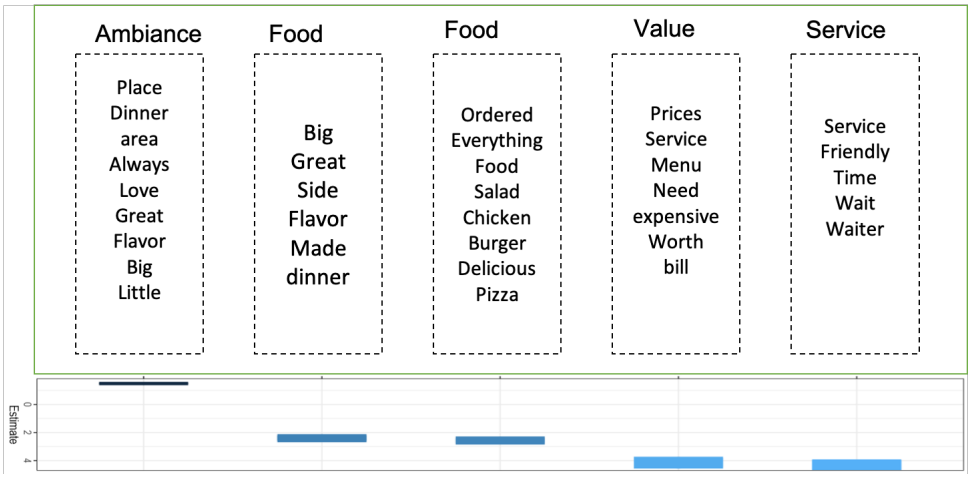
Rating : 4/5
Great place for an Italian dinner. We started with some house pasta and ceaser salad which were outstanding. One of the best pastas I have had. They have an outstanding dessert menu- I would not forget the chocolate cake with gelato ice-cream!The atmosphere is classy but relaxed. The service is quite decent given the amount of people they have every day. Parking can be painful but it was a Friday night and the whole city is out partying, should have Ubered!! I found the prices reasonable though the desserts were a bit expensive

Rating : 4/5
The look and feel of this Italian diner is amazing! Wow!! Such beautiful décor and such a large and comfortable seating area. The tables are large and the dinner area is quite airy. The food at most is ordinary, though certainly not bad - I am a big fan of Italian food and have had better pastas and ceasar salad. The prices are on the higher side but cmon this is a special place! There is a parking lot but it fills up too fast and Saturday nights can be a parking nightmare! The servers are nice people, kind and attentive but nothing extra ordinary

(b) Control Group

Figure 4.2: LDA and SLDA topics

(a) LDA topics for Yelp review corpus (with seed words)



(b) SLDA topics for Yelp review corpus

sharing and sparse connections. Weight-sharing means using the same filter to interact with different parts of the data and sparse connection refers to the fact that there are fewer links between the neurons in adjacent layers. These two features reduce the parameter space of the model to a great extent thereby lowering the training time and number of training examples needed. Thus, CNN-based models take relatively little time to train compared to fully-connected networks or sequential networks. Training a CNN involves fixing the weight matrix of the shared filters by repeatedly updating the weights with the objective of minimizing a loss function that captures how far the predicted classification of the model is from the true class of training data.

An embedded sentence vector of dimension $n \times d$ enters the convolution layer. Filters of height h (where filter height denotes length of n-gram captured) and width d act on the input vector to generate one feature map each. For illustration purposes, let us consider a filter matrix F of size $h \times d$ that moves across the entire range of the input I of size $n \times d$, convolving with a subset of the input of size $h \times d$ to generate a feature map M of dimension $(n-h+1) \times 1$. A typical convolution operation involves computing a map by element-wise multiplication of a window of word vectors with the filter matrix in the following manner:

$$M(i,1) = \sum_{i=1}^{n-h+1} \sum_{m=1}^{h} \sum_{n=1}^{d} I(i+(m-1),n)F(m,n) \tag{4.1}$$

When there is a combination of filters of varying heights (say 1,2,3 etc.), we get feature maps of variable sizes ($n, n-1, n-2$ and so on).

Max-pooling and flattening operations are performed to concatenate variable size feature maps into a single feature vector that is passed to the next feature generating layer.

The role of the convolutional layer in this model is to extract phrase-level location invariant features that can aid in attribute and sentiment classification. A feature map emerging from a convolution of word vectors can be visualized as several higher-order

representations of the original sentence like n-grams that capture negation like "not good" or "not that great experience" or n-grams that describe an attribute like "waiting staff" or "owner's wife." The number of filters to be used, N_f is fixed during hyper parameter tuning. Feature maps from all filters are passed through a non-linear activation function a_f with a small bias or constant term b to generate an output that would serve as input for the next stages of the model.

$$O_i = a_f(M_i + b) \tag{4.2}$$

The function f here can be any non-linear transformation that acts on the element-wise multiplication of the filter weights and word vectors plus a small bias term b. We use Rectified Linear Units (RELU) that is more robust in ensuring the network continues to learn for longer time periods compared to other activation functions like the tanh function (Nair and Hinton 2010). This activation function has the following format:

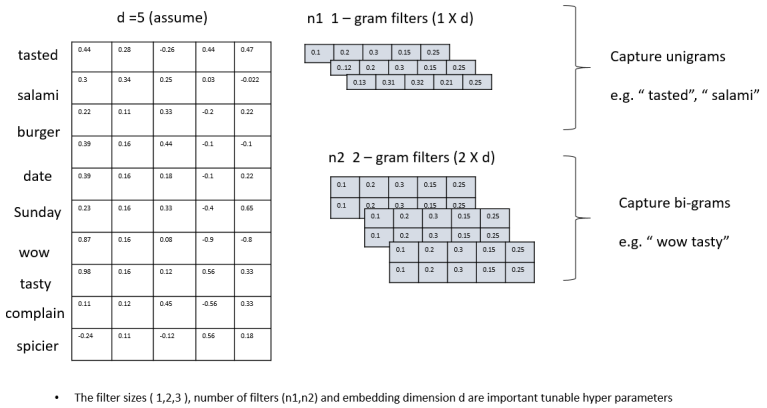
$$RELU(x) = max(0, x) \tag{4.3}$$

This activation function sets all negative terms in the feature maps to zero while preserving the positive outputs.

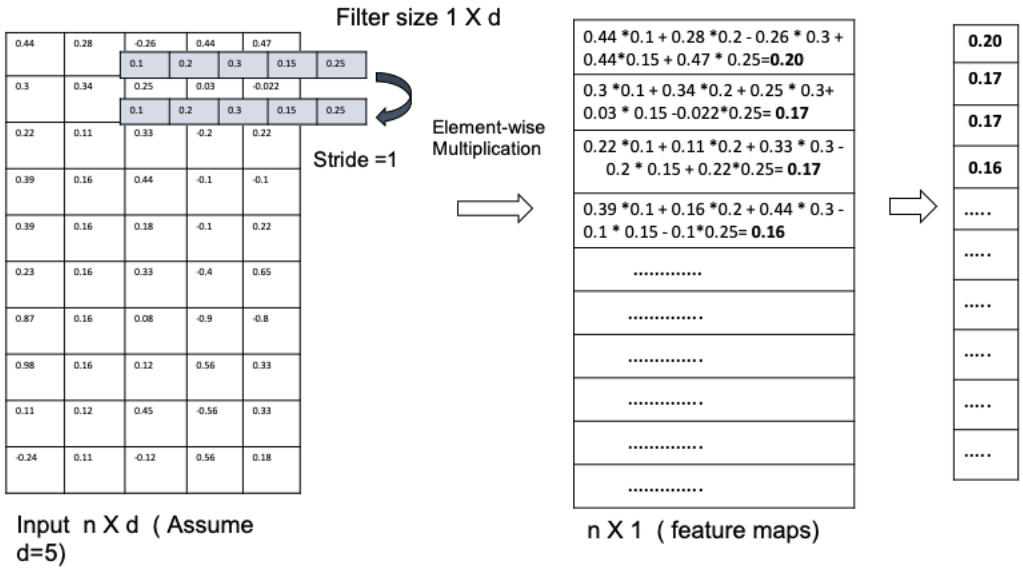
Figure 4.3 shows the structure of the convolution layer, the convolution operation respectively and a sample visualization of a feature map. During the course of training, each filter specializes in identifying a particular class. For instance, the filter in Figure 4.3 (c) has specialized in detecting *good food*.

Long Short Term Memory (LSTM) layer. The concatenated feature maps from the convolution layer are next fed into a Long Short Term Memory (LSTM) layer. LSTM is a special variant of the recurrent neural networks (RNN) that specialize in handling long-range dependencies. RNNs have a sequential structure and hence they can model inter-dependencies between the current input and the previous inputs using a history variable

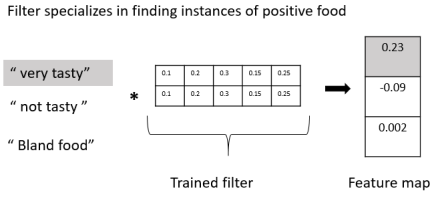
Figure 4.3: Convolutional Neural Network



(a) CNN filters and hyper parameters



(b) Convolution Operation

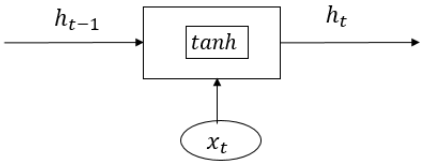


(c) Visualization of a feature map

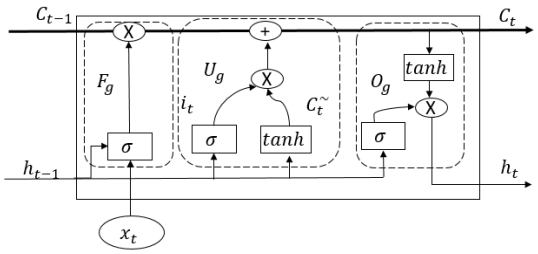
that is passed from one time period to the next. However, in practice, RNNs fail to do text classification tasks better than CNNs due to the "vanishing gradient" problem which causes a network to totally stop learning after some iterations (Nair and Hinton 2010). Vanishing gradients in the earlier layers of a recurrent neural network mainly result from a combination of non-linear activation functions like sigmoid and small weights in the later layers. LSTMs solve this problem by using a special memory unit with a fixed weight self-connection and linear activation function that ensures a constant non-vanishing error flow within the cell. Further, to ensure that irrelevant units do not perturb this cell, they employ a combination of gate structures that constantly make choices about what parts of the history need to be forgotten and what needs to be retained to improve the accuracy of the task at hand (Hochreiter and Schmidhuber 1997). This architecture has shown remarkable success in several natural language processing tasks like machine translation and speech to text transcription.

Figure 4.4 is a comparison of RNN and LSTM architectures. In an RNN, the output at a particular time t is fed back into the same network in a feedback loop. In this way, a new input x_t interacts with the old history variable h_{t-1} to create the new output o_t and the a new history variable h_t. This is like in a relay race where each cell of the network passes on information of its past state to the next cell (but each cell is identical, and therefore it is equivalent to passing on the information to itself). The Long Short Term Memory (LSTM) cell differs from the RNN cell on two important aspects—the existence of a cell state C_t (the long term memory) and a combination of gates that regulate the flow of information into the cell state. The cell state is like a conveyor belt that stores the information that the network decides to take forward at any point in time t. Gates are sigmoidal units whose value is multiplied with the values of the other nodes. If the gate has a value of zero, it can completely block the information coming from another node whereas if the gate has a value $\in (0,1)$, it can selectively allow some portion of the information to pass. Thus, gates are like "regulators" of what information flows into and remains active within the system.

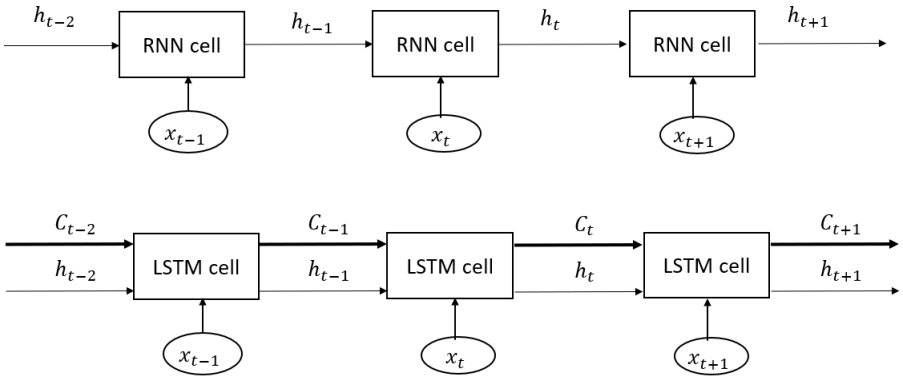
Figure 4.4: Comparison of RNN and LSTM cells (Goodfellow et al. 2016)



(a) RNN cell



(b) LSTM cell



(c) Unrolled RNN and LSTM networks

The LSTM has three gates — a forget gate G_F, an update gate G_U and an output gate G_O.

Suppose x_t represents the input to the LSTM at a particular time t and h_{t-1} denotes the hidden state (or history) that is stored from a previous time period. At the first stage, the forget gate decides what part of the previous state needs to be forgotten or removed from the cell state. For instance, in a long sentence, once the LSTM has figured out that the sentence is primarily about the taste of a burger, it might chose to remove useless information regarding weather or day of the week that says nothing about food taste. The transition function for the forget gate can be represented as :

$$f_t = \sigma(W_f \dot{[}h_{t-1}, x_t] + b_f) \tag{4.4}$$

This equation is a typical neural network equation that involves an element-wise multiplication of a weight function with the hidden state h_{t-1} and current input x_t followed by the addition of a bias term and subsequent non-linearity. The other transition functions of the LSTM include an update function and an output function. The update function decides what part of the current input needs to be updated to the cell state. The output function first determines the output o_t for the current time period and subsequently, the new hidden state h_t that is passed to the next time period by selectively combining the current output and cell state contents that seem most relevant.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{4.5}$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{4.6}$$

$$C_t = (f_t C_{t-1} + i_t \tilde{C}_t) \tag{4.7}$$

$$o_t = \sigma(W_o[h_t - 1, x_t] + b_o) \tag{4.8}$$

$$h_t = o_t \tanh(C_t) \tag{4.9}$$

All the weight matrices W_f, W_i, W_c and W_o are shared across different time steps. Thus, training an LSTM basically involves training these shared weight matrices by optimizing over a loss function.

4.2.1 Additional Performance Metrics

In this section, we describe some additional performance metrics that were excluded in the main text for brevity. These include some objective metrics like precision and recall (derived from the confusion matrices) and some practical considerations like model building effort, scalability and interpretability.

Precision and Recall See tables 4.2 and 4.3 to see how different models perform on metrics like precision, recall and F1 score.

Table 4.2: Precision and Recall (Sentiment Classification)

	CNN			CNN-LSTM (self-trained)			CNN-LSTM (Glove 300)		
Class	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Very Negative	38%	22%	28%	43%	24%	31%	30%	28%	29%
Negative	51%	27%	35%	52%	38%	44%	44%	59%	51%
Neutral	34%	37%	35%	46%	33%	39%	36%	18%	24%
Positive	34%	60%	43%	37%	68%	48%	40%	63%	49%
Very Positive	52%	40%	45%	67%	44%	53%	80%	58%	67%

Table 4.3: Precision and Recall (Attribute Analysis)

	CNN-LSTM (self trained)			CNN-LSTM (Glove 100)		
Class	Precision	Recall	F1	Precision	Recall	F1
ambiance	60%	58%	59%	55%	77%	64%
food	83%	79%	81%	86%	75%	80%
location	57%	56%	56%	73%	31%	43%
service	80%	60%	69%	71%	76%	73%
price	72%	75%	74%	76%	75%	76%

Model Building Time Lexicon models take approximately 175-180 hours of construction time. Most of the time is spent on human-tagging of the 8575 attribute and sentiment words into specific classes using Amazon's Mechanical Turk. Similarly, the creation of training and test data sets for the supervised learning algorithms takes approximately 100

hours.A human tagger takes around 1 minute to classify every word and 2-3 minutes to classify full sentences However, once created, we could use the same dataset to train and test a variety of machine learning and deep learning classifiers (e.g., SVM, Random Forest, Naive Bayes, CNN, LSTM and CNN-LSTM). After generating the training data, supervised learning models (including the deep learning models) need time for hyper parameter tuning and model training. Though this is an iterative process, all deep learning models take less than 10 minutes (in a quad core processor) for completing one training cycle and hence model calibration can be completed in 6-7 hours. Thus, model building is time-consuming for all algorithms but is a one-time activity.

Scalability The more time-sensitive metric is scalability i.e. the time required for a trained model to classify new examples. With respect to the scalability metric, the deep learning classifiers clearly outperform the lexicon based classifiers with the machine learning classifiers in between the other two. The main reason is the "look-up" method employed by lexicon based methods. Every word in a sentence needs to be sequentially searched through the entire lexicon to determine its class. Hence, the lexicon methods need several hours to classify our corpus of 27,332 reviews comprising of 999,885 sentences. On the other hand, deep learning models are able to classify our entire review dataset comprising in approximately 18- 20 minutes.

Interpretability refers to how well a machine classifier can explain the reasoning or logic behind its classifications (Doshi-Velez and Kim 2017). In general, text mining methods differ in their strengths and weakness across various dimensions, there is no one method that is superior in all dimensions. Though the CNN-LSTM model outperforms all the other models in accuracy and scalability, however, it falls short in terms of interpretability with respect to lexicon methods.

Chapter 5

Appendix to Chapter 2

5.1 Appendix: Data

5.1.1 Top National Chains (2017)

Table 5.1 presents some key details about the nationally established chains we study in our analysis. All data is for the year 2017. Negative WOM stands for share of negative reviews (1-2 star reviews), PWOM is the share of positive reviews (4-5 stars) and Neutral is share of 3-star reviews. The table shows that while there are some chains like McDonald's that have both lower ACSI scores as well as higher proportion of negative reviews, most other chains like Subway, Domino's Pizza, Papa John's and Pizza Hut have a very high proportion of negative reviews inspite of having good ACSI scores. The regression results in Table 2.2 remain similar if we exclude McDonald's. The chain dummy in that case is -0.91 and significant.

Table 5.1: Revenue, Satisfaction and Review Valence

Name	Stores (US)	Estd	Revenue (USD bn)	ACSI score	Brand (USD mn)	Star (Avg)	Negative WOM	Positive WOM	Neutral
Subway	25908	1965	11.3	80	18766	2.76	48%	38%	13%
McDonald's	14027	1955	37.64	69	126044	2.07	68%	20%	13%
Starbucks	13930	1971	17.65	78	44503	3.2	32%	50%	18%
Dunkin'	12538	1950	8.46	78	NA	2.6	54%	33%	13%
Pizza Hut	7522	1958	5.51	80	7372	2.19	66%	25%	9%
Burger King	7226	1953	9.65	76	6555	2.16	65%	20%	14%
Taco Bell	6446	1962	9.79	74	5213	2.64	53%	36%	12%
Wendy's	5769	1969	9.31	77	NA	2.29	62%	25%	13%
Domino's Pizza	5587	1960	5.93	79	7446	2.63	54%	37%	9%
KFC	4109	1952	NA	77	15131	1.78	77%	13%	9%
Arby's	3415	1964	3.63	79	NA	2.84	46%	40%	14%
Papa John's	3314	1984	1.78	80	NA	2.38	61%	30%	9%
Chipotle	2364	1993	4.48	79	4422	3.03	41%	46%	13%
Chick-fil-A	2100	1946	9	87	NA	3.74	23%	66%	11%

5.1.2 Dictionary for expectation and novelty words

Table 5.2 is the dictionary we used to count occurrences of compare and novel words.

Table 5.2: Custom Dictionaries

Expect	Novel	Employee
anticipate	curiosity	back office
belief	curious	bartender
brand	fresh	boy
change	innovative	desk
changed	learn	employee
consistent	new	front desk
contrary	novel	girl
declined	now	reception
deteriorate	offbeat	receptionist
exceed	recent	staff
expect	surprised	waiter
expectation	unique	waitress
expected	unusual	wait-staff
image	weird	
improve		
improved		
inconsistent		
met		
notion		
past		
prior		
recall		
remember		
reputation		
standard		
standards		
unexpected		
worsen		

5.2 Appendix: Proofs

5.2.1 Proof of Lemma 2.4.1

Let price p be such that F has no mass point at the belief thresholds $\underline{\phi}(p)$, $\overline{\phi}(p)$, $\tilde{\phi}(p;(\mu_h,\mu_\ell))$ for $\mu_h,\mu_\ell \in \{0,1\}$.

In a *full WOM equilibrium*, a follower purchases after an $\emptyset$-message iff $\phi \geq \Phi(p)$. It exists iff an early adopter wants to talk after both $q = \ell, h$, i.e., exactly iff followers are uninformed.

Analogously, a *no WOM equilibrium* exists iff an early adopter does not want to talk regardless of his signal, i.e., exactly iff followers are well-informed.

In a *negative WOM equilibrium*, $\tilde{\phi}(p;(0,1)) \leq \Phi(p)$ is decreasing in β. It exists iff an early adopter does not want to talk with $q = h$, and wants to talk otherwise, i.e.,

$$\beta \in B^{\text{neg}}(p) := \left\{ \beta \in [0,1] | F(\tilde{\phi}(p;(0,1))) - F(\underline{\phi}(p)) \leq \frac{\xi}{1-\beta} \leq F(\bar{\phi}(p)) - F(\tilde{\phi}(p;(0,1))) \right\}.$$

For optimistic followers, $B^{\text{neg}}(p) = [0,1]$, so negative WOM equilibria always exists.

If followers are not optimistic, since $\lim_{\beta \to 0} \tilde{\phi}(p,(0,1)) = \Phi(p)$ and $\lim_{\beta \to 0} \frac{\xi}{1-\beta} = \xi$, there exists a threshold $\hat{\beta}^{\text{neg}}(p) > 0$ such that $\beta \leq \hat{\beta}^{\text{neg}}(p) \implies \beta \notin B^{\text{neg}}(p)$. So, for $\beta \leq \hat{\beta}^{\text{neg}}(p)$, negative WOM equilibria cannot exist.

Analogously, in a *positive WOM equilibrium*, $\tilde{\phi}(p;(1,0)) \geq \Phi(p)$ is increasing in β and it exists iff

$$\beta \in B^{\text{pos}}(p) := \left\{ \beta \in [0,1] | F(\bar{\phi}(p)) - F(\tilde{\phi}(p;(1,0))) \leq \frac{\xi}{1-\beta} \leq F(\tilde{\phi}(p;(1,0))) - F(\underline{\phi}(p)) \right\}.$$

Hence, a positive WOM equilibrium always exists if followers are pessimistic and does not exists if $\beta \leq \hat{\beta}^{\text{pos}}(p)$ for $\hat{\beta}^{\text{pos}}(p) > 0$.

5.2.2 Proof of Proposition 2.4.2

The demand function depends on the WOM equilibrium played. In a full WOM equilibrium, the demand, defined as the probability of a follower buying, is given by

$$D^{\text{full}}(p) = (1-\beta)(1-F(\Phi(p))) + \beta\ (\pi(\phi_0)(1-F(\underline{\phi}(p))) + (1-\pi(\phi_0))\ (1-F(\bar{\phi}(p)))).$$

With no WOM,

$$D^{\text{no}}(p) = 1 - F(\Phi(p)).$$

With negative WOM,

$$D^{\text{neg}}(p) = (1-\beta+\beta\pi(\phi_0))\,(1-F(\tilde{\phi}(p;(0,1))))+\beta(1-\pi(\phi_0))(1-F(\bar{\phi}(p))).$$

With positive WOM,

$$D^{\text{pos}}(p) = (1-\beta+\beta\,(1-\pi(\phi_0)))(1-F(\tilde{\phi}(p;(1,0))))+\beta\pi(\phi_0)(1-F(\underline{\phi}(p))).$$

The price p determines whether followers are optimistic, pessimistic, well-informed, or uninformed, and Lemma 1 then pins down the type of WOM when no threshold is ϕ_0, which we analyse separately. We compute cutoffs $p^{\text{neg}}, p^{\text{pos}}$ such that $p < p^{\text{neg}} \Leftrightarrow \beta \in B^{\text{neg}}(p)$ and $p < p^{\text{pos}} \Leftrightarrow \beta \in B^{\text{pos}}(p)$.

1. If $\phi_0 < \underline{\phi}(p)$, then $F(\underline{\phi}(p)) = F(\bar{\phi}(p)) = F(\Phi(p)) = 1$. Followers are well-informed. No WOM is the unique equilibrium. Since profits are zero, the firm never induces this case.

2. If $\underline{\phi}(p) < \phi_0 < \Phi(p)$, then $F(\underline{\phi}(p)) = 0 < F(\Phi(p)) = F(\bar{\phi}(p)) = 1$. Followers are pessimistic. A positive WOM equilibrium exists for all $\beta \in [0,1]$. Negative WOM arises iff $F(\tilde{\phi}(p;(0,1))) = 0$ or $\tilde{\phi}(p;(0,1)) = \phi_0$ which is satisfied iff

   $$\tilde{\phi}(p;(0,1)) \leq \phi_0 \quad \Leftrightarrow \quad \beta \geq \frac{\Phi(p)-\phi_0}{\Phi(p)-\phi_0+(1-\Phi(p))\phi_0\pi_H-\Phi(p)(1-\phi_0)\pi_L} =: \hat{\beta}^{\text{neg}}(p).$$

   Otherwise, $F(\tilde{\phi}(p;(0,1))) = 1$. Hence, $B^{\text{neg}}(p) = \left(\hat{\beta}^{\text{neg}}(p), 1-\xi\right]$ and $\hat{\beta}^{\text{neg}}(p) > 0$ if $\Phi(p) > \phi_0$ because the denominator is strictly positive.

   Finally, let p^{pess} be such that $p < p^{\text{pess}} \Leftrightarrow \underline{\phi}(p) < \phi_0$. It is defined implicitly by $\phi_0 = \underline{\phi}(p^{\text{pess}})$ or $p^{\text{pess}} = \frac{\phi_0\pi_H^2+(1-\phi_0)\pi_L^2}{\phi_0\pi_H+(1-\phi_0)\pi_L}$. $\hat{\beta}^{\text{neg}}(p)$ is increasing in p and hence, $\beta \geq \hat{\beta}^{\text{neg}}(p)$

iff

$$p \leq p^{\text{neg}} := \frac{\phi_0(\pi_H - \pi_L)(1 - \beta(1 - (\pi_H + \pi_L))) - \beta(1 - \pi_L)\pi_L + \pi_L}{1 - \beta(1 - (\phi_0 \pi_H + (1 - \phi_0)\pi_L))}.$$

One can show that for all $\beta > 0$, $p^{\text{pess}} > p^{\text{neg}}$, so at any $p < p^{\text{neg}}$ priors are pessimistic, but a negative WOM equilibrium can exist.

3. If $\phi_0 = \underline{\phi}(p)$, either a no or positive WOM equilibrium is played as 2.

4. If $\Phi(p) < \phi_0 < \bar{\phi}(p)$, then $F(\underline{\phi}(p)) = F(\Phi(p)) = 0 < F(\bar{\phi}(p)) = 1$, i.e., followers are optimistic. Full or no WOM equilibrium cannot exist. A negative WOM equilibrium exists for all $\beta \in [0,1]$. Positive WOM exists iff $F(\tilde{\phi}(p;(1,0))) = 1$ or $\tilde{\phi}(p;(1,0)) = \phi_0$ which is satisfied iff

$$\begin{aligned} \tilde{\phi}(p;(1,0)) \geq \phi_0 \quad &\Leftrightarrow \quad \beta \geq \frac{\phi_0 - \Phi(p)}{\phi_0 - \Phi(p) + \Phi(p)(1-\phi_0)(1-\pi_L) - (1-\Phi(p))\phi_0(1-\pi_H)} \\ &=: \hat{\beta}^{\text{pos}}(p) \end{aligned}$$

Otherwise, $F(\tilde{\phi}(p;(1,0))) = 0$. $\hat{\beta}^{\text{pos}}(p) > 0$ for $\phi_0 > \Phi(p)$ for the same reason as $\hat{\beta}^{\text{neg}}(p) > 0$ when followers are pessimistic. Hence, $B^{\text{pos}}(p) = \left[\hat{\beta}^{\text{pos}}(p), 1 - \xi\right]$.

Finally, let p^{opt} be such that $p < p^{\text{opt}} \Leftrightarrow \Phi(p) < \phi_0$. It is defined implicitly by $\phi_0 = \Phi(p^{\text{opt}})$ or $p^{\text{opt}} = \phi_0 \pi_H + (1 - \phi_0)\pi_L$. $\hat{\beta}^{\text{pos}}(p)$ is decreasing in p and hence, $\beta \geq \hat{\beta}^{\text{pos}}(p)$ iff

$$p \geq p^{\text{pos}} := \frac{\phi_0(\pi_H - \pi_L)(\beta(\pi_H + \pi_L) - 1) + \pi_L(\beta \pi_L - 1)}{\beta(\phi_0(\pi_H - \pi_L) + \pi_L) - 1}$$

One can show that for all $\beta \in (0,1]$, $p^{\text{pos}} < p^{\text{opt}}$, i.e., if $p < p^{\text{opt}}$, both positive and negative WOM equilibria exist. Further, $p^{\text{pess}} > p^{\text{opt}}$ because $x \mapsto x^2$ is convex and $p^{\text{neg}} > p^{\text{opt}}$, i.e., the firm always sets the price as close as possible to p^{neg} (p^{pess}) to induce negative (positive) WOM where followers are pessimistic.

5. If $\phi_0 = \Phi(p)$, both positive and negative WOM equilibria exist for all $\beta > 0$, since $\tilde{\phi}(p;(0,1)) < \phi_0 < \tilde{\phi}(p;(1,0))$ for $\beta > 0$.

6. If $\phi_0 > \bar{\phi}(p)$, then $F(\underline{\phi}(p)) = F(\Phi(p)) = F(\bar{\phi}(p)) = 0$. Followers are well-informed. No WOM is the unique equilibrium. Such beliefs are induced iff

$$p < p^{\text{well}} := \frac{\pi(\phi_0) - (\phi_0 \pi_H^2 + (1-\phi_0)\pi_L^2)}{1 - \pi(\phi_0)}$$

7. If $\phi_0 = \bar{\phi}(p)$, then either no one talks as in 6, or negative WOM arises as in 4.

Lastly, we compare profits. In a positive WOM equilibrium given p, profits are

$$\Pi^{\text{pos}}(p) := p\, D^{\text{pos}}(p) = p\, \beta \pi(\phi_0).$$

It is then straightforward to compute the maximal profit when positive WOM is induced is

$$\Pi^{\text{pos}}(p^{\text{pess}}) \quad = \beta(\phi_0 \pi_H^2 + (1-\phi_0)\pi_L^2).$$

Analogously, with negative WOM the maximum profit is

$$\begin{aligned} \Pi^{\text{neg}}(p^{\text{neg}}) \quad &= (1-\beta)\pi(\phi_0) + \beta(\phi_0 \pi_H^2 + (1-\phi_0)\pi_L^2)) \\ &> \Pi^{\text{pos}}(p^{\text{pess}}) \qquad \text{for } \beta < 1. \end{aligned}$$

The maximal profit with no WOM is $\Pi^{\text{no}}(p^{\text{well}}) = p^{\text{well}}$. Negative WOM is an equilibrium iff $\Pi^{\text{neg}}(p^{\text{neg}}) \geq \Pi^{\text{no}}(p^{\text{well}}) \Leftrightarrow \beta \leq \bar{\beta}^{\text{hom}}$, and no WOM iff $\beta \geq \bar{\beta}^{\text{hom}}$. No other WOM equilibria can exist.

First, define for any $p \in \mathscr{P}^*$, $\overline{\xi}(p) := F(\overline{\phi}(p)) - F(\Phi(p))$ and $\underline{\xi}(p) := F(\Phi(p)) - F(\underline{\phi}(p))$. Further, for any given $\beta > 0$ and $w \in \{\text{full}, \text{no}, \text{neg}, \text{pos}\}$, define

$$\hat{\mathscr{P}}^w(\beta) := \arg\max_p D^w(p)p.$$

Note that as $\beta \to 0$, $\hat{\mathscr{P}}^w(\beta)$ converges to $\mathscr{P}^*$.

1. With a strong brand image, it follows immediately that $\xi > \underline{\xi}(p)$ for all $p \in \mathscr{P}^*$. Further, for any $\varepsilon > 0$, there exists a $\bar{\beta}$ so that for all $\beta < \bar{\beta}$ and $\hat{p} \in \hat{\mathscr{P}}^w(\beta)$, there exists a $p^* \in \mathscr{P}^*$ so that $|p^* - \hat{p}| < \varepsilon$. Hence, for sufficiently small β,

$$\xi > F(\Phi(\hat{p})) - F(\underline{\phi}(\hat{p}))$$

   for all $\hat{p} \in \hat{\mathscr{P}}^w(\beta)$ with $w \in \{\text{full}, \text{no}, \text{neg}, \text{pos}\}$. Thus, there is no profit maximizing equilibrium price that induces positive WOM.

2. With a weak brand image and $\pi_L = 0$, we have that for all $p \in \mathscr{P}^*$, $\underline{\xi}(p) = F(\Phi(p)) > \xi$. Again, for any $\varepsilon > 0$, there exists a $\bar{\beta}$ so that for all $\beta < \bar{\beta}$ and $\hat{p} \in \hat{\mathscr{P}}^w(\beta)$, there exists a $p^* \in \mathscr{P}^*$ so that $|p^* - \hat{p}| < \varepsilon$. Thus, for sufficiently small β

$$\xi < F(\Phi(\hat{p})) - F(\underline{\phi}(\hat{p}))$$

   for all $\hat{p} \in \hat{\mathscr{P}}^w(\beta)$ with $w \in \{\text{full}, \text{no}, \text{neg}, \text{pos}\}$. Thus, in any PBE positive WOM must occur.

Given a WOM regime $w \in \{\text{full}, \text{no}, \text{neg}, \text{pos}\}$, the firm maximizes $\max_p pD^w(p)$. $D^w(p)$ uniformly converges to $1 - F(\Phi(p))$. Thus, for sufficiently small β, any profit-maximizing price is arbitrarily close to p^*. Denote an arbitrary sequence of solutions by $p^w(\beta)$. Let $\overline{\xi} := F(\overline{\phi}(p^*)) - F(\Phi(p^*))$ and $\underline{\xi} := F(\Phi(p^*)) - F(\overline{\phi}(p^*))$. If $\xi < \min\{\overline{\xi}, \underline{\xi}\}$, then for sufficiently small β

$$\xi < \min\{(1-\beta)(F(\overline{\phi}(p^w(\beta))) - F(\Phi(p^w(\beta)))), (1-\beta)(F(\Phi(p^w(\beta))) - F(\underline{\phi}(p^w(\beta))))\},$$

so that followers are well-informed, and Lemma 2.4.1 implies the unique equilibrium features no WOM.

The other three cases $\xi > \max\{\overline{\xi}, \underline{\xi}\}$, $\overline{\xi} > \xi > \underline{\xi}$, $\overline{\xi} > \xi > \underline{\xi}$, follow analogously.

Proof of Proposition 2.5.2 We characterize the pooling equilibrium in which the H-type firm maximizes its profits. Because the price is not informative on equilibrium path, beliefs of the followers and early adopters are the same on equilibrium path whether or not the firm has private information. Hence, Lemma 2.4.1 applies unchanged. However, the induced demand functions faced by an θ-type firm are now different, as the firm knows that it is a θ type. Thus, for an equilibrium price p^*, the demand faced by a θ-type firm in a no, full, negative, and positive WOM equilibrium, respectively is:

$$\begin{aligned}
D^{\text{no}}(p;\theta) &= 1 - F(\Phi(p)) \\
D^{\text{full}}(p;\theta) &= \beta\ (\pi_\theta(1 - F(\underline{\phi}(p))) + (1-\pi_\theta)\ (1 - F(\bar{\phi}(p)))) \\
&\quad + (1-\beta)\ (1 - F(\Phi(p))), \\
D^{\text{neg}}(p;\theta) &= (1 - \beta + \beta\pi_\theta)\ (1 - F(\tilde{\phi}(p;(0,1)))) + \beta(1-\pi_\theta)(1 - F(\bar{\phi}(p))) \\
D^{\text{pos}}(p;\theta) &= (1 - \beta + \beta\ (1-\pi_\theta))(1 - F(\tilde{\phi}(p;(1,0)))) + \beta\pi_\theta(1 - F(\underline{\phi}(p))).
\end{aligned}$$

1. We start with the case of homogeneous priors $F = 1(\phi \geq \phi_0)$. Since the WOM stage is identical to the baseline case, points 1-7 of the proof of Proposition 2.4.2 apply here too. However, the profits are different. The maximal profit of and H-type firm when positive WOM is induced is given by

$$\Pi^{\text{pos}}(p^{\text{pess}};H) \quad := p^{\text{pess}}\ \beta\pi_H = \tfrac{\phi_0\pi_H^2 + (1-\phi_0)\pi_L^2}{\pi(\phi_0)}\beta\pi_H.$$

The maximal profit of an H-type firm when negative WOM is induced is given by

$$\begin{aligned}
\Pi^{\text{neg}}(p^{\text{neg}};H) &:= p^{\text{neg}}\ (1 - \beta + \beta\pi_H)) \\
&= \tfrac{\phi_0(\pi_H - \pi_L)(1 - \beta(1 - (\pi_H + \pi_L))) - \beta(1-\pi_L)\pi_L + \pi_L}{(1 - \beta + \beta\pi(\phi_0))}\ (1 - \beta + \beta\pi_H) \\
&> \Pi^{\text{pos}}(p^{\text{pess}})
\end{aligned}$$

for sufficiently small β. The maximal profit with no WOM is given by

$$\Pi^{\text{no}}(p^{\text{well}};H) \quad := \frac{\pi(\phi_0)-(\phi_0\pi_H^2+(1-\phi_0)\pi_L^2)}{1-\pi(\phi_0)} < \Pi^{\text{neg}}(p^{\text{neg}};H)$$

for sufficiently small β. Hence, for sufficiently small $\beta > 0$, there can only be negative WOM. No other WOM equilibria can be sustained.

2. For the uniform distribution, F is the identity function. Again an analogous argument to the proof of Proposition 2.4.2 can be applied with the adjusted demand and profit functions. Note, that the profit-maximizing price as β tends to zero is, however, identical and equal to $\frac{\pi_H}{2}$. Hence, the exact same proof can be applied.

Bibliography

Aggarwal CC, Zhai C (2012) *Mining text data* (Springer Science & Business Media).

Angelis MD, Bonezzi A, Peluso AM, Rucker DD, Costabile M (2012) On braggarts and gossips: A self-enhancement account of word-of-mouth generation and transmission. *Journal of Marketing Research* 49(4):551–563.

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management science* 57(8):1485–1509.

Athey S, Bayati M, Doudchenko N, Imbens G, Khosravi K (2018) Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research.

Balaji P, Nagaraju O, Haritha D (2017) Levels of sentiment analysis and its challenges: A literature review. *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 436–439 (IEEE).

Bass FM (1969) A new product growth for model consumer durables. *Management science* 15(5):215–227.

Bengio Y, CA M (2015) Rmsprop and equilibrated adaptive learning rates for nonconvex optimization. *corr abs/1502.04390* .

Berger J (2014) Word of mouth and interpersonal communication: A review and directions for future research. *Journal of consumer psychology* 24(4):586–607.

Berger J, Sorensen AT, Rasmussen SJ (2010) Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science* 29(5):815–827.

Biyalogorsky E, Gerstner E, Libai B (2001) Customer referral management: Optimal reward programs. *Marketing Science* 20(1):82–95.

Brahma S (2018) Improved sentence modeling using suffix bidirectional lstm. *arXiv preprint arXiv:1805.07340* .

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

Campbell A (2013) Word-of-mouth communication and percolation in social networks. *American Economic Review* 103(6):2466–98.

Campbell A, Mayzlin D, Shin J (2017) Managing buzz. *The RAND Journal of Economics* 48(1):203–229.

Carman A (2018) Why do you leave restaurant reviews? *The Verge* URL `https://www.theverge.com/2018/4/10/17215784/whyd-you-push-that-button-yelp-tripadvisor-reviews-restaurant`, accessed July 27, 2020.

Chakraborty I, Kim M, Sudhir K (2019) Attribute sentiment scoring with online text reviews: Accounting for language structure and attribute self-selection. *Cowles Foundation Discussion Paper* .

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43(3):345–354.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science* 29(5):944–957.

Churchill GA, Collins RH, Strang WA (1975) Should retail salespersons be similar to their customers. *Journal of Retailing* 51(3):29.

Dhar V, Chang EA (2009) Does chatter matter? the impact of user-generated content on music sales. *Journal of Interactive Marketing* 23(4):300–307.

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

Duan W, Gu B, Whinston AB (2008) Do online reviews matter?—an empirical investigation of panel data. *Decision support systems* 45(4):1007–1016.

Evans FB (1963) Selling as a dyadic relationship–a new approach. *American Behavioral Scientist* 6(9):76–79.

Feldman R (2013) Techniques and applications for sentiment analysis. *Communications of the ACM* 56(4):82–89.

Fradkin A, Grewal E, Holtz D, Pearson M (2015) Bias and reciprocity in online reviews: Evidence from field experiments on airbnb. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 641–641 (ACM).

Frazier GL, Summers JO (1984) Interfirm influence strategies and their application within distribution channels. *Journal of Marketing* 48(3):43–55.

Galeotti A (2010) Talking, searching, and pricing*. *International Economic Review* 51(4):1159–1174.

Galeotti A, Goyal S (2009) Influencing the influencers: a theory of strategic diffusion. *The RAND Journal of Economics* 40(3):509–532.

Ganu G, Elhadad N, Marian A (2009) Beyond the stars: improving rating predictions using review text content. *WebDB*, volume 9, 1–6 (Citeseer).

Ghose A, Ipeirotis PG (2007) Designing novel review ranking systems: predicting the usefulness and impact of reviews. *Proceedings of the ninth international conference on Electronic commerce*, 303–310 (ACM).

Gilchrist DS, Sands EG (2016) Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy* 124(5):1339–1382.

Godes D (2016) Product policy in markets with word-of-mouth communication. *Management Science* 63(1):267–278.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing science* 23(4):545–560.

Godes D, Mayzlin D, Chen Y, Das S, Dellarocas C, Pfeiffer B, Libai B, Sen S, Shi M, Verlegh P (2005) The firm's management of social interactions. *Marketing Letters* 16(3):415–428.

Gonçalves P, Araújo M, Benevenuto F, Cha M (2013) Comparing and combining sentiment analysis methods. *Proceedings of the first ACM conference on Online social networks*, 27–38.

Goodfellow IJ, Bengio Y, Courville AC (2016) *Deep Learning*. Adaptive computation and machine learning (MIT Press).

Grice HP, Cole P, Morgan J, et al. (1975) Logic and conversation. *1975* 41–58.

Gurney N, Loewenstein G (2019) Filling in the blanks: What restaurant patrons assume about missing sanitation inspection grades. *Journal of Public Policy & Marketing* 0743915619875419.

Harris ZS (1954) Distributional structure. *Word* 10(2-3):146–162.

He S, Hollenbeck B, Proserpio D (2020) The market for fake reviews. *Available at SSRN* .

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput.* 9(8):1735–1780, ISSN 0899-7667.

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

Huang JL, Liu M, Bowling NA (2015) Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology* 100(3):828.

Jin G, Lee J, Luca M, et al. (2018) Aggregation of consumer ratings: an application to yelp. com. *Quantitative Marketing and Economics* 16(3):289–339.

Joachims T (2002) *Learning to classify text using support vector machines*, volume 668 (Springer Science & Business Media).

Johnson R, Zhang T (2017) Deep pyramid convolutional neural networks for text categorization. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol 1: Long Papers)*.

Jones SE, LeBaron CD (2002) Research on the relationship between verbal and nonverbal communication: Emerging integrations. *Journal of communication* 52(3):499–521.

Kamada Y, Öry A (2017) Contracting with word-of-mouth management. *Cowles Foundation Discussion Paper No. 2048* .

Ke TT, Shin J, Yu J (2020) A model of brand positioning: Product-portfolio view. *working paper* .

Keramatfar A, Amirkhani H (2019) Bibliometrics of sentiment analysis literature. *Journal of Information Science* 45(1):3–15.

Kim Y (2014) Convolutional neural networks for sentence classification. *arXiv:1408.5882* .

Kornish LJ, Li Q (2010) Optimal referral bonuses with asymmetric information: Firm-offered and interpersonal incentives. *Marketing Science* 29(1):108–121.

Kotler P (2000) *Marketing Management: The Millennium Edition* (Prentice Hall, Inc.).

Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology* 5(3):213–236.

Leduc MV, Jackson MO, Johari R (2017) Pricing and referrals in diffusion on networks. *Games and Economic Behavior* 104:568–594.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *Journal of Marketing Research* 48(5):881–894.

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

Li Y, Lu SF, Lu LX (2019) Do yelp reviews influence consumer choice in the presence of government ratings? evidence from us nursing homes. *Evidence from US Nursing Homes (October 1, 2019)* .

Little RJ, Rubin DB (2019) *Statistical analysis with missing data*, volume 793 (John Wiley & Sons).

Liu B, et al. (2010) Sentiment analysis and subjectivity. *Handbook of natural language processing* 2(2010):627–666.

Liu X, Lee D, 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.

Longpre L, Durmus E, Cardie C (2019) Persuasion of the undecided: Language vs. the listener. *Proceedings of the 6th Workshop on Argument Mining*, 167–176.

Luca M (2016) Reviews, reputation, and revenue: The case of yelp. com .

Luca M, Vats S (2013) Digitizing doctor demand: The impact of online reviews on doctor choice. *Cambridge, MA: Harvard Business School* .

Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science* 62(12):3412–3427.

Luo X (2009) Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. *Marketing Science* 28(1):148–165.

Manzoor E, Chen GH, Lee D, Smith MD (2020) Influence via ethos: On the persuasive power of reputation in deliberation online. *arXiv preprint arXiv:2006.00707* .

Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* 104(8):2421–55.

Mayzlin D, Shin J (2011) Uninformative advertising as an invitation to search. *Marketing Science* 30(4):666–685.

Mehrabian A (2017) *Nonverbal communication* (Routledge).

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Mittal V, Katrichis JM, Kumar P (2001) Attribute performance and customer satisfaction over time: evidence from two field studies. *Journal of Services Marketing* .

Mittal V, Kumar P, Tsiros M (1999) Attribute-level performance, satisfaction, and behavioral intentions over time: a consumption-system approach. *Journal of Marketing* 63(2):88–101.

Mittal V, Ross Jr WT, Baldasare PM (1998) The asymmetric impact of negative and positive attribute-level performance on overall satisfaction and repurchase intentions. *The Journal of Marketing* 33–47.

Mohan K, Pearl J (2018) Graphical models for processing missing data. *arXiv preprint arXiv:1801.03583* .

Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 807–814, ICML'10, ISBN 978-1-60558-907-7.

Nanli Z, Ping Z, Weiguo L, Meng C (2012) Sentiment analysis: A literature review. *2012 International Symposium on Management of Technology (ISMOT)*, 572–576 (IEEE).

Nelson P (1970) Information and consumer behavior. *Journal of Political Economy* 78(2):311–329, ISSN 00223808, 1537534X, URL `http://www.jstor.org/stable/1830691`.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Science* 31(3):521–543.

Nielsen FÅ (2011) A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903* .

Nosko C, Tadelis S (2015) The limits of reputation in platform markets: An empirical analysis and field experiment. Technical report, National Bureau of Economic Research.

Onishi H, Manchanda P (2012) Marketing activity, blogging and sales. *International Journal of Research in Marketing* 29(3):221–234.

Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, volume 10.

Peloza J, Ye C, Montford WJ (2015) When companies do good, are their products good for you? how corporate social responsibility creates a health halo. *Journal of Public Policy & Marketing* 34(1):19–31.

Pennebaker JW (1997) Writing about emotional experiences as a therapeutic process. *Psychological science* 8(3):162–166.

Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

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

Qazi A, Raj RG, Hardaker G, Standing C (2017) A systematic literature review on opinion types and sentiment analysis techniques. *Internet Research* .

Rubin DB (1976) Inference and missing data. *Biometrika* 63(3):581–592.

Scherer KR, London H, Wolf JJ (1973) The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality* 7(1):31–44.

Schoenmüller V, Netzer O, Stahl F (forthcoming) The extreme distribution of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research* .

Schouten K, Frasincar F (2015) Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 28(3):813–830.

Sebastiani F (2002) Machine learning in automated text categorization. *ACM computing surveys (CSUR)* 34(1):1–47.

Seiler S, Yao S, Wang W (2017) Does online word of mouth increase demand?(and how?) evidence from a natural experiment. *Marketing Science* .

Sheth J (1976) Buyer-seller interaction: A conceptual framework. *Advances in Consumer Rasearch* 3(3B):382—386.

Shmueli-Scheuer M, Herzig J, Konopnicki D, Sandbank T (2019) Detecting persuasive arguments based on author-reader personality traits and their interaction. *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 211–215.

Slovic P, MacPhillamy D (1974) Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior and Human Performance* 11(2):172–194.

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Spiro RL, Weitz BA (1990) Adaptive selling: Conceptualization, measurement, and nomological validity. *Journal of marketing Research* 27(1):61–69.

Sun M (2012) How does the variance of product ratings matter? *Management Science* 58(4):696–707.

Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267–307.

Timoshenko A, Hauser JR (2018) Identifying customer needs from user-generated content. *Marketing Science (Forthcoming)* .

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research* 51(4):463–479.

Van Zant AB, Berger J (2020) How the voice persuades. *Journal of personality and social psychology* 118(4):661.

Wang H, Lu Y, Zhai C (2010) Latent aspect rating analysis on review text data: a rating regression approach. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 783–792 (ACM).

Wang J, Yu LC, Lai KR, Zhang X (2016) Dimensional sentiment analysis using a regional cnn-lstm model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol 2: Short Papers)*, 225–230.

Wang XS, Lu S, Li X, Khamitov M, Bendle N (2021) Audio mining: The role of vocal tone in persuasion. *Journal of Consumer Research* .

Williams KC, Spiro RL (1985) Communication style in the salesperson-customer dyad. *Journal of marketing Research* 22(4):434–442.

Wilson DT (1976) Dyadic interaction: an exchange process. *ACR North American Advances* .

Xu X (2019) Examining the relevance of online customer textual reviews on hotels' product and service attributes. *Journal of Hospitality & Tourism Research* 43(1):141–163.

Zhou C, Sun C, Liu Z, Lau F (2015) A c-lstm neural network for text classification. *preprint arXiv:1511.08630* .

Zhu F, 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.

ProQuest Number: 28321251

INFORMATION TO ALL USERS
The quality and completeness of this reproduction is dependent on the quality
and completeness of the copy made available to ProQuest.

Distributed by ProQuest LLC (2021).
Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license
or other rights statement, as indicated in the copyright statement or in the metadata
associated with this work. Unless otherwise specified in the copyright statement
or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization
of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA