

COMPARING LATENT DIRICHLET ALLOCATION AND LATENT
SEMANTIC ANALYSIS AS CLASSIFIERS

Leticia H. Anaya, B.S., M.S.

Dissertation Prepared for the Degree of
DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2011

APPROVED:

Nicholas Evangelopoulos, Major Professor
Shailesh Kulkarni, Committee Member
Robert Pavur, Committee Member
Dan Peak, Committee Member
Nourredine Boubekri, Committee Member
Mary C. Jones, Chair of the Department of
Information Technology and
Decision Sciences
Finley Graves, Dean of College of Business
James D. Meernik, Acting Dean of the
Toulouse Graduate School

Anaya, Leticia H. Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers. Doctor of Philosophy (Management Science), December 2011, 226 pp., 40 tables, 23 illustrations, references, 72 titles.

In the Information Age, a proliferation of unstructured text electronic documents exists. Processing these documents by humans is a daunting task as humans have limited cognitive abilities for processing large volumes of documents that can often be extremely lengthy. To address this problem, text data computer algorithms are being developed.

Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are two text data computer algorithms that have received much attention individually in the text data literature for topic extraction studies but not for document classification nor for comparison studies. Since classification is considered an important human function and has been studied in the areas of cognitive science and information science, in this dissertation a research study was performed to compare LDA, LSA and humans as document classifiers. The research questions posed in this study are:

- R1: How accurate is LDA and LSA in classifying documents in a corpus of textual data over a known set of topics?
- R2: How accurate are humans in performing the same classification task?
- R3: How does LDA classification performance compare to LSA classification performance?

To address these questions, a classification study involving human subjects was designed where humans were asked to generate and classify documents (customer comments) at two levels of abstraction for a quality assurance setting. Then two computer algorithms, LSA and LDA, were used to perform classification on these documents.

The results indicate that humans outperformed all computer algorithms and had an accuracy rate of 94% at the higher level of abstraction and 76% at the lower level of abstraction. At the high level of abstraction, the accuracy rates were 84% for both LSA and LDA and at the lower level, the accuracy rate were 67% for LSA and 64% for LDA.

The findings of this research have many strong implications for the improvement of information systems that process unstructured text. Document classifiers have many potential applications in many fields (e.g., fraud detection, information retrieval, national security, and customer management). Development and refinement of algorithms that classify text is a fruitful area of ongoing research and this dissertation contributes to this area.

Copyright 2011

by

Leticia H. Anaya

ACKNOWLEDGEMENTS

I want to express my deepest gratitude and appreciation to my dissertation committee. To my Chair, Dr. Nicholas Evangelopoulos, for his mentoring, encouragement and guidance in my pursuit of the PhD. I honestly can never thank him enough for opening my eyes into a field that I never knew existed, for sharing his knowledge and for his overall support in this endeavor. To Dr. Robert Pavur who taught me how to think analytical way beyond my expectations. To Dr. Shailesh Kulkarni who showed me the beauty of incorporating analytical mathematics for manufacturing operations which I hope that one day I can expand upon. To Dr. Dan Peak, whose encouragement and words of wisdom were always welcomed and I thank him for sharing those interesting and funny anecdotes, stories, and images that made life interesting while pursuing this endeavor. To Dr. Nourredine Boubekri for his support and for his initial words of encouragement that allowed me to overcome the fear that prevented me from even taking the first step toward the pursuit of a PhD. To Dr. Victor Prybutok for his support, his encouragement, his sense of humor, his sense of fairness and overall great disposition in spite of having dealt with much stronger trials and tribulations that I ever had. To Dr. Sherry Ryan and Dr. Mary Jones, two of my female professors who left an everlasting positive impression in my life not only in terms of their scholastic work but also in terms of their attitudes toward life. To my family and all my PhD fellow friends who shared the PhD experience with me.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	iii
LIST OF TABLES	viii
LIST OF FIGURES	x
1. OVERVIEW	1
1.1 Introduction.....	1
1.2 Overview of Document Classification.....	2
1.3 Modeling Document Classification	3
1.4 Overview of Human Classification.....	4
1.5 Overview of Latent Semantic Analysis	5
1.6 Overview of Latent Dirichlet Allocation	6
1.7 Research Question and Potential Contributions	7
2. BASIC TEXT MINING CONCEPTS AND METHODS	10
2.1 Introduction.....	10
2.2 Text Mining Introduction.....	10
2.3 Vector Space Model.....	12
2.4 Text Mining Methods	14
2.4.1 Latent Semantic Analysis	14
2.4.2 Probabilistic Latent Semantic Analysis	17
2.4.3 Relationship between LSA and pLSA	18
2.4.4 The Latent Dirichlet Allocation Method.....	20
2.4.5 Gibbs Sampling: Approximation to LDA.....	23

2.5 Conclusion	25
3. CLASSIFICATION	25
3.1 Classification as Supervised Learning	26
3.2 Data Points Representation	27
3.3 What are Classifiers?	28
3.4 Developing a Classifier	28
3.4.1 Loss Functions	30
3.4.2 Data Selection	31
3.5 Classifier Performance	32
3.5.1 Classification Matrix	32
3.5.2 Receiver Operating Characteristics (ROC) Curves	33
3.5.3 F1 Micro-Sign and F1 Macro-Sign Tests	34
3.5.4 Misclassification Costs	35
3.6 Document Classification	36
3.7 Commonly Used Document Classifiers	37
3.7.1 K-Nearest Neighborhood Classifier	37
3.7.2 Centroid-Based Classifier	38
3.7.3 Support Vector Machine	38
3.8 Conclusion	39
4. HUMAN CATEGORIZATION LITERATURE REVIEW	41
4.1 Introduction	41
4.2 Categorization at the Developmental Stages	42
4.3 Category Learning Theories	44

4.4 Human Categorization Modeling.....	45
4.5 Categorization Issue: Base Rates (Prior Probabilites)	46
4.6 Categorization Issue: Inverse Base Rate Effect (IBRE)	47
4.7 Recent Trends	50
4.8 Conclusion	53
5. THEORY DEVELOPMENT	54
5.1 Introduction.....	54
5.2 Latent Semantic Analysis (LSA) as a Proposed Theory of Knowledge.....	54
5.3 Latent Dirichlet Allocation as a Proposed Model of Human Cognition.....	55
5.4 Prototype and Exemplar Theories.....	56
5.5 Linking the Prototype and Exemplar Theory.....	58
5.6 Generic Mixture of Classification.....	59
6. RESEARCH METHODOLOGY.....	63
6.1 Overview.....	63
6.2 First Stage: Comment Generation.....	63
6.3 Second Stage: Human Categorization	66
6.4 Management Interpretation of Comments	67
6.5 Classification Performance Measures.....	68
6.6. Analysis of First Set of Surveys	69
6.7 Human Classification Experiment	74
6.7.1 Analysis of Second Set of Surveys	75
6.7.2 Human Categorization Results	76
6.8. LDA Categorization Procedure.....	78

6.8.1 LDA Topic Verification.....	81
6.8.2 LDA Results.....	87
6.9 LSA Results	92
6.10 Discussion.....	97
7. FUTURE RESEARCH DIRECTION.....	103
APPENDIX A. MODELING OF HUMAN CATEGORIZATION	112
APPENDIX B. ONLINE SURVEY FOR PART I OF RESEARCH.....	121
APPENDIX C. ONLINE SURVEY FOR PART II OF RESEARCH.....	134
APPENDIX D. IRB10480 INFORMATION NOTICE.....	192
APPENDIX E. NOVEMBER 8, 2010 IRB APPLICATION	195
APPENDIX F IRB10504 INFORMATION NOTICE	203
APPENDIX G . NOVEMBER-30-2010, IRB APPLICATION.....	206
APPENDIX H. LDA ALGORITHM INFORMATION.....	214
REFERENCES	219

LIST OF TABLES

	Page
2.1 Two Document Sample.....	12
3.1 Training Loss Functions and Regularizers for 8 Classifiers (Li and Yang, 2003)	31
3.2 Classification Matrix.....	32
3.3 Sample Classification Matrix.....	35
3.4 Cost Classification Matrix	36
6.1 First Surveys Inter-Rater Probabilities.....	67
6.2 First Surveys Subtopic Classification Scores	69
6.3 First Surveys Main Categories Classification Scores	70
6.4 ANOVA of Main Categories	72
6.5 95% Confidence Intervals for Main Categories.....	72
6.6 Tukey Method Grouping of Main Categories.....	72
6.7 ANOVA of Subtopics.....	73
6.8 95% Confidence Intervals for Subtopics	74
6.9 Second Survey Set Inter-Rater Reliabilities.....	75
6.10 Human Classification Subtopic Classification Scores	76
6.11 Unique Human Classification Subtopics	77
6.12 Main Categories Human Classification Scores.....	78
6.13 LDA Extracted Topics	81
6.14 LDA Extracted Documents.....	82
6.15 LDA Topic 1 Verification.....	82
6.16 LDA Topic 2 Verification.....	83

6.17 LDA Topic 3 Verification.....	84
6.18 LDA Topic 4.Verification.....	85
6.19 Topic Designation.....	87
6.20 Frequency Distribution of Main Categories	87
6.21 Topic Designation-Trial 1.....	88
6.22 LDA Main Categories Classification Scores	88
6.23 LDA Main Categories Macro-F1 Scores	89
6.24 LDA Main Categories Summary Classification Scores.....	89
6.25 LDA Subtopic Summary Classification Scores.....	90
6.26 LDA Subtopic Detailed Classification Scores.....	90
6.27 Unique LDA Categorization Subtopics	92
6.28 LSA-IDF Main Categories Classification Scores	92
6.29 LSA-TF Main Categories Classification Scores.....	92
6.30 LSA Subtopic Summary Classification Scores.....	93
6.31 LSA Specific Subtopic Classification Scores	94
6.32 Unique LSA-TF Categorization Subtopics.....	96
6.33 Unique LSA-IDF Categorization Subtopics	96
6.34 Summary of Unique Categorization Subtopics for all Methods.....	98
6.35 Overall Summary of Results.....	99

LIST OF FIGURES

	Page
2.1 Documentation representation in vector space	13
2.2 SVD of original X matrix	15
2.3 SVD of truncated X matrix	15
2.4 Loading on words matrix	16
2.5 Loading on documents matrix	16
2.6 Plate notation of pLSA model.....	17
2.7 Decomposition of PLSA into matrices	19
2.8 Plate notation of the LDA model	20
2.9 Decomposition of LDA into matrices	21
3.1 Classifier with 41% misclassification Rate.....	30
3.2 Classifier with 29% misclassification Rate.....	30
3.3 Support vector machine	39
5.1 Prototype theory model.....	57
5.2 Exemplar theory model.....	57
5.3 General mixture classification example.....	60
6.1 Fish-bone diagram	64
7.1 The attention learning covering map (ALCOVEL) model	107
7.2 Artificial neural network.....	108
A.1 The component-cue model.....	114
A.2 The attention distinctive input (ADIT) model	115
A.4 Casual modeling of features in a stimulus	118
I-1. Basic LDA Flowchart.....	218

CHAPTER 1

OVERVIEW

1.1 Introduction

The Information Age has created a proliferation of documents that is constantly increasing while our ability to process and absorb this data has remained constant (Feldman and Sanger, 2007). With an “information” explosion being a fact of the 21st century, electronic documents are being produced, stored, and accessed at exponential rates (Lavengood and Kiser, 2007). The Internet has been the biggest driving force behind this proliferation of documents and its ever increasing expansion has had a bigger impact on the information field than any other communication media ever developed. According to Rodgers (1998), “It took radio 38 years to reach 50 million listeners. It took television 13 years to get the same effect. But the Internet took 4 years to reach the same number and traffic is doubling every 100 days.” This ever expanding network continues to have a big impact on the volumes of electronics documents that are constantly being generated. Spira (2006) mentioned in his article that it takes now the world about 15 minutes to create a quantity of digital information equivalent to that currently in the Library of Congress and much less is also created through non-digital means. This document proliferation has created a situation where we have to incorporate modern technology to be able to process and handle the massive amounts of electronic documents that are being generated on a daily basis.

Although the Internet is a major force in generating unstructured text (e.g., blogs, wikis, e-mails, chat-rooms, online surveys, etc.), businesses are also generating unstructured text data by streamlining their traditional document management practices to online operations. For many

businesses, the mission seems to be to eradicate paper by streamlining their record management practices to maintaining electronic records of each transaction (Anonymous, 2011). Thus, unstructured text is constantly being produced, stored, and accessed through various means and it is generally being placed at our disposal through the click of a computer mouse. Processing these documents by humans is a daunting task not only because of their large quantities but also because the length of many of these documents can be extremely long. Processing of these documents usually requires our cognitive abilities to perform content analysis, classification and deduce topics. However, as humans, we cannot possibly perform these tasks on the ever increasing volume of unstructured text that is being generated. To address this problem, a number of algorithms for text data processing are being developed.

In this research (dissertation) I focus on classification as an important cognitive function that humans usually perform on a daily basis. I design a study where I ask humans to classify documents and then I assess how two partially automated methods, latent dirichlet allocation (LDA) and latent semantic analysis (LSA), perform classification on similar documents. I compare the performance of these methods to human classification and to each other. To begin, a brief overview of document classification is presented in this chapter, followed by overviews on human categorization, LSA and LDA. Lastly, the research questions for this dissertation are presented.

1.2 Overview of Document Classification

Document classification is a data mining technique that is performed on documents. A document could be an abstract, a web page, an e-mail, a comment, or any corpus of words that requires classification. Document classification differs from the traditional data mining classification in three different aspects: the data structure, the vector representation of an

instance, and the variable representation. The data structure in traditional data mining classification consists of instances or tuples (rows of a database table) that represent structured data with variables taking many different formats and which can be qualitative as well as quantitative. In document classification, an instance (or document), however, is represented as an unstructured “bag of words.” The words in the document are the variables and the manner in which these words are ordered is irrelevant. In traditional data mining, the vector representation of the instance in dimensional space is much smaller than the vector representation of the words of a document in document classification--which can range in orders of magnitude much greater than in traditional data mining. Lastly, in traditional data mining, the variables themselves can have many different representation, numerical, nominal, etc. But in document text mining, the frequency of a word appearance is the only value assigned to each word variable.

1.3 Modeling Document Classification.

Two types of document classification exist in the literature: content classification and collective classification. Content document classification deals with classifying documents that are not linked to each other in any way. Thus, each document is mutually exclusive from another document. In content classification, part of the data, the training set, is usually used to derive a classifier f . If the classifier, f , needs to be redefined, another part of the data, the validation set, is created to modify the classifier. Once a classifier is deemed suitable, its effectiveness is tested on the remaining data, the testing set.

Collective classification deals with classifying documents that are linked to each other by some variable. The linking variables could be authors, conference, citation, web page links, etc. Namata et al. (2009) model document collective classification as a network of larger nodes $V = \{V_1, \dots, V_n\}$ where each larger node represent a document connected via links. The set of

documents V is further subdivided into two sets of nodes: X , the observed values nodes (labels and words) and Y , (labels) the nodes whose values need to be determined. The goal in collective classification is to infer labels for the unlabeled nodes from the existing documents. The problem of collective classification becomes one of optimization in which a set of neighboring nodes N_i , usually a fully labeled training set, such that $N_i \subseteq V$ are used to describe the underlying network structure or used to derive the classifier and use it to predict a new label for an unlabeled document. The classification task is to label the nodes $Y_i \in Y$ with a label y_i from a set of labels $L = \{L_1, \dots, L_q\}$. A classifier f takes as input the values of neighboring nodes N_i as arguments and returns a label value for Y_i from the label set L . The classifiers that can be used in collective document classification can be similar to the ones used in content document classification.

However, in collective document classification one issue that could exist is that the training data set N_i may be incomplete or not fully labeled. For these types of situations the iterative classification algorithm explained by Namata et al. (2009) can be used. Here, the iterative classification problem takes a classifier f as input the values of N_i as arguments and returns a label value for Y_i from the label set L . Since not all the values in N_i are known, the predicted Y_i values used iteratively to relabel the next Y_i value until all the labels stabilize.

1.4 Overview of Human Classification

The ability for humans to categorize is considered to be a foundation of cognition in human categorization literature (Arterberry and Bornstein, 2002). The human categorization field is vast as human categorization has been studied in the literature from many different aspects. Human categorization studies range from simulation modeling of the mechanisms of the human categorization mental process to childhood cognition developmental studies. Appendix A

provides an overview of some of the human categorization simulation models have been created and evolved over time. The human categorization field has consisted of two areas: category learning and category representation. Recently, the paradigm for the field of human categorization has focused, both empirically and theoretical, on two main issues: a) the field embracing the cognitive neuroscience field and b) the human category learning processes being mediated by multiple distinct learning systems instead of a single learning system (Ashby and Maddox, 2005). This paradigm shift has been affected the two traditional areas of human categorization: category learning and category representation. These two areas are now linked to mechanisms in the brain. According to Ashby and Maddox (2005), the brain's neural mechanism that mediate the learning of different categories are different from the brain's neural structures that mediate the representation of learned categories. As evidence of the existence of this difference, Ashby and Maddox (2005) mention that frontal damaged patients and Parkinson's disease patients that are impaired in category learning can still perform category representation tasks by recognizing old familiar categories (e.g., animals, fruits, etc.) and no evidence exists that impairment in category representation (or failing to recognizing old familiar categories or development of agnosias) can lead to general learning category problems.

1.5 Overview of Latent Semantic Analysis

Text mining methods have been developed to analyze all structured and unstructured text data. Latent semantic analysis (LSA) is a text mining method that is used in informational retrieval of documents and it is an improvement over the traditional vector space model of documents. In the vector space model, when a document is represented as vector of words in vector space, a document can be retrieved by comparing its similarity to the original document (Salton, 1995). The dot product of two vectors can be used to compute the angle between two

vectors and then this angle can be used to determine how similar one document is to another document. However, the vector space model of documents does not address the issue of synonymy or polysemy that is an inherent part of documents. Synonymy refers to the grouping of different words with identical and similar meanings (e.g., the following list of words: “cribs,” “toys,” and “pacifiers” belong to the concept or topic of “babies”). Polysemy refers to words that can have more than one meaning. “Java” can refer to a computer programming language, but it also can refer to “coffee.”

Latent semantic analysis (LSA) was developed in an effort to resolve some of the polysemy and synonymy issues found in document retrieval. LSA is a least square algorithm where a document is a mixture of words and a corpus is a collection of documents LSA works by reducing an original term-by-document matrix representation of a corpus of documents into a filtered term-document matrix through a procedure called singular value decomposition (SVD). When a smaller document representation of words in vector space is compared to another vector of words, the belief is that it is generally easier to compare two smaller vectors for similarity purposes (and thus facilitate the retrieval of documents) than to compare two larger vectors. The input to the LSA algorithm is generally a term-by-document matrix \mathbf{X} and the output is a filtered term-by-document matrix $\tilde{\mathbf{X}}$.

Since LSA associates documents to topical factors, it can be used for document classification. Chapter 2 provides some technical details on LSA implementation, as well as using LSA as a classifier.

1.6 Overview of Latent Dirichlet Allocation

Latent dirichlet allocation (LDA) is a Bayesian method for topic extraction in a collection of documents. LDA is an extension of the probabilistic latent semantic analysis (pLSA) method-

a model that was considered to be an improvement over LSA. As with pLSA, the latent dirichlet allocation method is based on the probabilistic topic model, the aspect model. LDA was introduced in Blei et al. (2003). According to the basic LDA model, each word in each document is selected from a mixture of topic distributions. A document consists of a distribution of topics. LDA estimates the parameters in the word and document topic distributions using Markov chain Monte Carlo (MCMC) simulations. In this study, the implementation of LDA is achieved through the implementation of Griffiths and Steyvers (2004) Gibbs sampling algorithm.

1.7 Research Questions and Potential Contribution

LDA and LSA have been used extensively for topic extraction in a set of corpus of documents. They, however, have not been used as classifiers of documents nor their performance has been compared to the actual performance of human subjects. The research questions posed in this study are:

RQ1: How accurate are LDA and LSA in classifying documents in a corpus of textual data over a known set of topics?

RQ2: How accurate are humans in classifying over a known set of topics?

RQ3: How do LDA, LSA, and human classification performances compare to each other?

The contributions of this research have many strong implications for the improvement of information systems that process text data. The LSA and LDA algorithms covered in this dissertation can be used not only to evaluate documents stored in such information systems but also to facilitate the retrieval of such documents. By measuring the algorithms performance in document classification, it can be determined how effective these algorithms (compared to traditional used algorithms) will be for such information systems. By categorizing documents

effectively through classification algorithms, spam e-mails can be deleted before they reach the user, pornographic websites can be prevented from being shown to minors, terroristic web sites can be identified earlier, etc. For document retrieval, classification algorithms are also important components of the process. Classification algorithms are first used to identify the documents that are relevant to a query, and then once identified, these documents are ranked by their level of importance prior to retrieving them. When a Google Internet search is performed, the first task of the Google search engine is to identify relevant documents that match that query. This task is usually accomplished through a classification algorithm that classifies each web page as being relevant to the query or not. The second task is accomplished through an algorithm called PageRank that evaluates the importance of each relevant web page based on the number of incoming and outgoing links associated with that web page. This process ensures that the final links of the web pages displayed are listed in order of importance and relevance to the initial query. No knowledge exists as to whether Google has considered LDA as a possible algorithm for information retrieval.

This study is important as classification algorithms are needed to process the increased amount of documentation generated through the Internet in the form of web pages and in the form of comments generated through social network sites that have expanded exponentially. Facebook publishes its statistics periodically and for the month of August, 2011, it claims that it has more than 750 million active users and people interact with more than 900 million objects that include web pages, groups, events, and community pages (Facebook 2011). LinkedIn as of August 2011, claims that it has more than 120 million users (LinkedIn 2011). These comments posted on web pages or discussed in groups can also be analyzed through LSA and LDA

algorithms. By performing either topic extraction or document classification on these comments, it can be determined whether these comments are a cause of concern to the general population.

The remainder of this dissertation is organized as follows: First, a literature review of basic text mining concepts and methods is provided. This includes detailed coverage of LSA and LDA. Second, a literature review on classification is provided that details popular classifiers, measures used, and issues that are relevant in the field of classification. Third, a literature review of on a subset of classification (human classification) is provided. Human classification is a vast field and an attempt is made here to provide the reader with enough knowledge about the issues that exist in the human classification field. Fourth, the theory development chapter seeks to establish the connection between the human classification and the classification algorithms (LSA and LDA) used in this study. Fifth, the methodology chapter explains the procedure for conducting this study and provides the associated results. The sixth chapter discusses the findings and implications for future research. Along with this, appendixes relevant to the dissertation are included. This includes an appendix on modeling of human classification, an accepted decision science conference paper arising from this research, and an appendix on the Steyvers' and Griffiths LDA (2004) algorithm.

CHAPTER 2

BASIC TEXT MINING CONCEPTS AND METHODS

2.1 Introduction

This chapter provides the background to the text mining methods that were used as part of this dissertation research. This chapter serves as an introduction to text mining concepts and to the theory behind the text mining methods that will be used: latent semantic analysis (LSA) and the latent dirichlet allocation (LDA). A description of each method is provided as well as the description of an intermediate method, probabilistic latent analysis (pLSA), which is the basis for the development of the LDA method. The LSA and LDA methods are implemented in the form of computer algorithms and they are the “tools” that were used to conduct the research in this dissertation.

2.2 Text Mining Introduction

Text mining is a special data mining process that is used increasingly to analyze unstructured text data with the main objective of being to make some sense out of it. The development of modern information systems that have been increasingly capturing and storing unstructured text data has increased the importance of text mining. Extensive text data is not only being stored, but queries into such data are also increasingly being done in the form of natural language as the objective of such queries is to have computers interact more effectively with humans. But such queries can also benefit from text mining methods as sometimes there is a difference between what a user requests and what the user actually wants. Thus, the analyses of stored text data and human-like queries through text mining methods can also be used to yield additional information on databases. In addition to querying into databases, text mining methods

can be used in other areas. Lee et al. (2010) mentioned that text mining methods are used for topic discovery, target advertising, customer relationship management and the identification of the intellectual core of information systems.

Data mining has been applied in many different settings. In the medical field, data mining can be used to analyze databases to determine whether a treatment can be applied to a certain population or not. Yeh et al. (2011) applied data mining methods to dialysis patients' biochemical data to develop a decision support system that would allow them to predict hospitalization of hemodialysis patients. In the financial world, data mining methods are used for bankruptcy prediction, credit card approval, loan decision, stock analysis and even fraud detection (Ravisankar et al., 2011). In the academic world, databases of student records also benefit from data mining methods. Practically, any setting that stores numeric and/or text data can benefit from data mining methods.

Text mining is used to analyze documents and the input to any text mining method is a document-term matrix. A document is an object that is defined by the text mining analyst. A document can be a sentence, a single paragraph, an essay, a webpage, an e-mail message, a chapter, a newspaper article, an abstract or practically any written text unit—the final unit choice chosen depends on the objective on the analysis. A “term” is usually a specially selected word that appears with a certain frequency in a document and is deemed to contribute a certain amount of informational value to the document. Thus, not all words in a document can be classified as a “term” for text mining analysis. Some words that do not add substantial information to a document and are usually deleted are words like “the,” “and,” “for,” etc. Regarding the terms found in a document, the “bag of words” concept is applied to text mining. This concept

indicates that the *order* of appearance of the terms in a document is irrelevant for text mining analysis.

2.3 Vector Space Model

The basic assumption in any text data mining method is that a document can be represented as a vector of terms in the vector space model (VSM) (Salton, 1975). The importance of a term is denoted by its frequency of appearance in the document. The distance between the two vectors can be used to determine the similarity between these two vectors. The determination of how closely any two vectors are to each other depends on the dot product between two vectors.

The dot product of two vectors **A** and **B** is given by:

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}||\mathbf{B}|\cos(\theta) \quad (1)$$

To illustrate geometrically, the following two simple documents are represented in vector space. In this example Document 1 consists of two words, Word 1 and Word 2, and each word appears 5 and 2 times respectively in Document 1. Similarly, Word 1 and Word 2 appear 3 and 6 times respectively in Document 2.

Table 2.1

Two Document Sample

	Word 1	Word 2
Doc 1	5	2
Doc 2	3	6

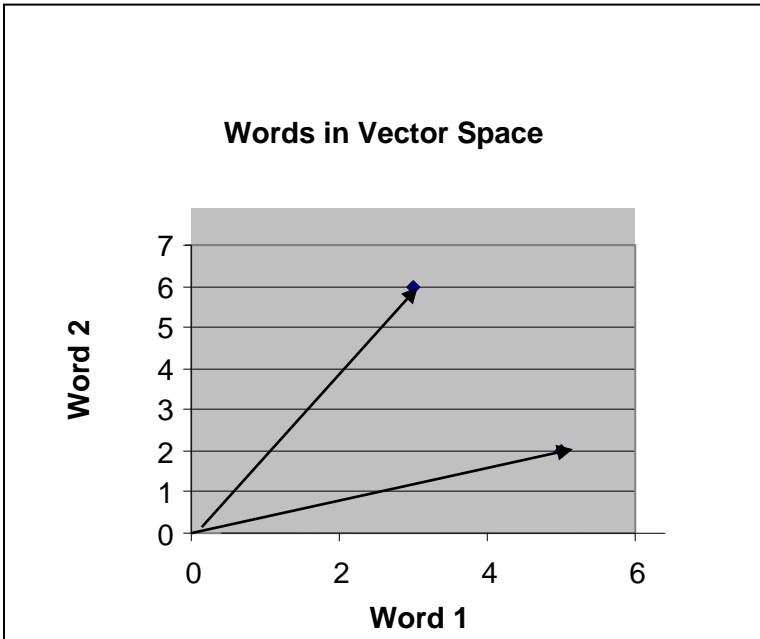


Figure 2.1. Document representation in vector space.

The cosine of the angle between two vectors can be used to measure the degree of similarity between two vectors. The angle between two vectors is used in information retrieval of documents as documents are retrieved based on the degree of similarity to another document. But comparing two documents using the dot product for information retrieval purposes is not without issues. One problematic issue is that it is very difficult to match similar documents that are very lengthy. The next issue is that it is unclear whether the order of words, which is usually ignored under the “bag of words” concept, may yield any informational value in text mining. Lastly, the issue of synonymy and polysemy are not addressed through the vector space model. Synonymy refers to the grouping of different words with identical and similar meanings (e.g., the following list of words: “cribs,” “toys,” and “pacifiers” belong to the concept or topic of “babies.”). Polysemy refers to words that can have more than one meaning. “Java” can refer to a computer programming language, but it also can refer to “coffee.” Thus, the sole

representation of documents as vectors of terms in vector space does not address polysemy nor synonymy issues. Using the vector space model, a collection of d documents can be represented in a space defined by a vocabulary of t terms by matrix \mathbf{X} , a $t \times d$ matrix containing the frequency of occurrence of each term in the vocabulary, in each document in the collection.

2.4 Text Mining Methods

Text mining methods have been developed to analyze all structured and unstructured text data. In this chapter, a brief description of the most well-known methods is given. These methods are: latent semantic analysis (LSA), the probabilistic latent semantic analysis pLSA, and the latent dirichlet allocation (LDA).

2.4.1 Latent Semantic Analysis

Latent semantic analysis (LSA) is a text mining method that was designed to resolve some of the polysemy and synonymy issues. LSA reduces the original term-document matrix \mathbf{X} into a filtered term-document matrix through the singular value decomposition (SVD) process. In LSA, a document representation as a vector of words in vector space can be compared to another vector for retrieval of similar documents. By reducing the document representation in the vector space, the retrieval of similar document has been made easier as it is easier to compare shorter length vectors for similarity purposes.

SVD decomposes the original matrix into three matrixes: a document eigenvector matrix, an eigenvalue matrix, and a term eigenvector matrix. The SVD of a rectangular matrix \mathbf{X} is given by:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (2)$$

where \mathbf{U} is the $t \times r$ matrix of eigenvectors of the square symmetric matrix of term covariances $\mathbf{X}\mathbf{X}^T$, \mathbf{V} is the $d \times r$ matrix of eigenvectors of the square symmetric matrix of document covariances $\mathbf{X}^T\mathbf{X}$ and $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix containing the square roots of eigenvalues (singular values) of both $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$, and $r \leq \min(t, d)$ is the rank of matrix \mathbf{X} .

Graphically, the LSA method can be illustrated as shown in Figure 2.2. The input to the LSA method is the \mathbf{X} matrix that is decomposed through singular value decomposition. The output to the LSA method is a truncated $\tilde{\mathbf{X}}$ matrix as shown in Figure 2.3.

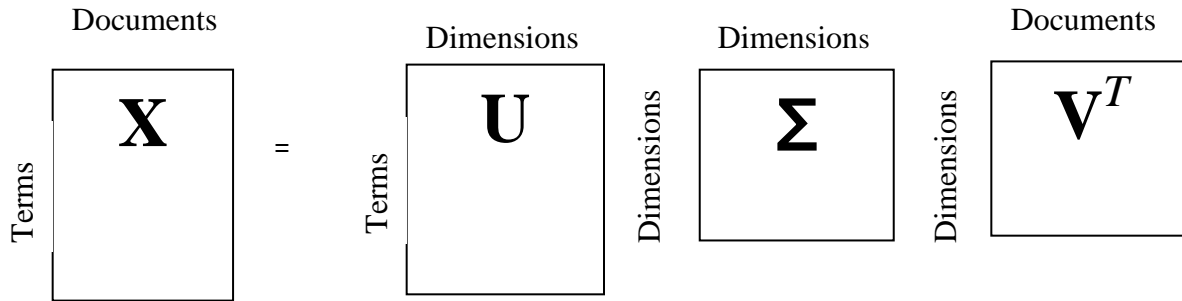


Figure 2.2 SVD of original \mathbf{X} matrix.

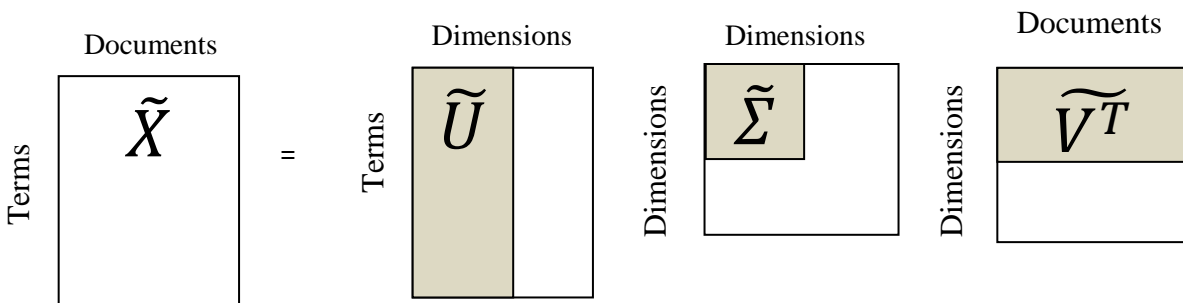


Figure 2.3 SVD of truncated $\tilde{\mathbf{X}}$ matrix.

In LSA, the original matrix \mathbf{X} can be approximated by another matrix $\tilde{\mathbf{X}}$ that results from truncating the middle eigenvalue matrix, $\mathbf{\Sigma}$, and multiplying it with the other associated matrices

that have been truncated to the same rank level as the middle matrix. Also, topics can be extracted by using factor loading and matrix rotation as shown in Figure 2.4 and Figure 2.5.

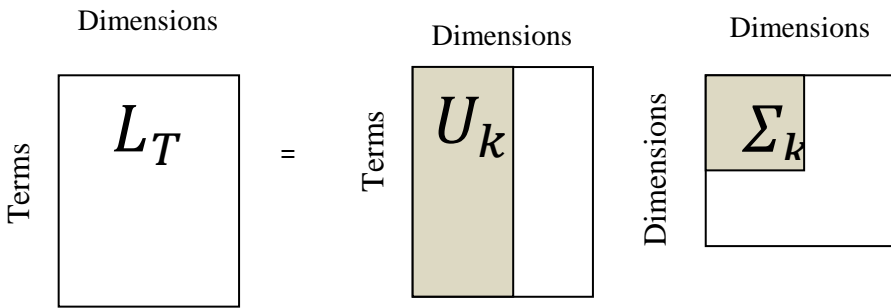


Figure 2.4. Loading on words matrix.

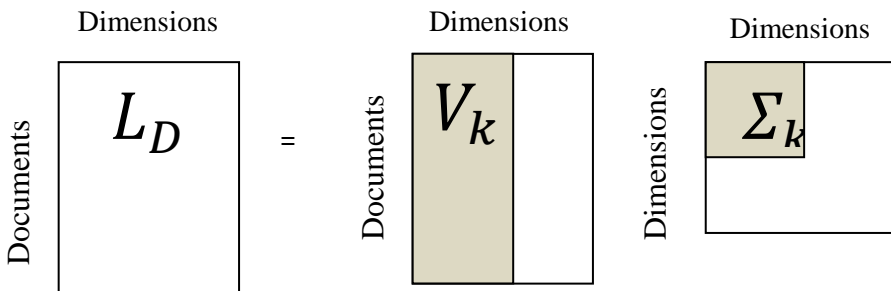


Figure 2.5. Loading on documents matrix.

When is it appropriate to apply LSA and what are the limitations of LSA? Papadimitriou et al. (2000) investigated three appropriate conditions to apply LSA. These conditions are: A) documents having the same writing style, B) each document being centered on a single topic and C) a word having a high probability of belonging to one topic but low probability of belonging to other topics. The limitations of LSA is that LSA is based on dimension reduction of the original dataset. The determination of dimension factors is based on a subjective judgment. Another limitation to LSA is that LSA has orthogonal characteristics and because of this, multiple

occurrences of a word from different factors (topics) is usually prevented. Because of the orthogonal characteristics of dimension factors, the words in a topic have a high relation with words in that topic but little with words in other topics. Thus, words like “Java” are likely to be found in one topic and not in another. Because of this orthogonality, LSA tends to prevent multiple occurrences of a word in different topics and thus LSA cannot be used effectively to resolve polysemy issues.

2.4.2 Probabilistic Latent Semantic Analysis (pLSA)

The next text mining technique that was developed to improve upon LSA was the probabilistic latent semantic analysis (pLSA) technique. pLSA is also known as probabilistic latent semantic indexing (pLSI). pLSA was introduced in 1999 by Thomas Hofman and it was designed to deal better with polysemy issues than LSA (Hofman, 2001). pLSA is an extension of LSA and it provides a statistical foundation to LSA. The model for pLSA is the statistical model called the aspect model. The assumptions for the aspect model is the “bag of words” assumption and the probability of a word being in a document $p(w,d)$ is generated independently. The model assumes that a document, d , is generated with a probability $p(d)$, and a topic, z , is picked with a probability $p(z/d)$ and a word, w , is generated with a probability $p(w/z)$. The pLSA models the probability of selecting a word in a document as:

$$p(w,d) = \sum_z p(z) p(w/z) p(d/z) \quad (3)$$

In plate notation, the pLSA model can be represented as:

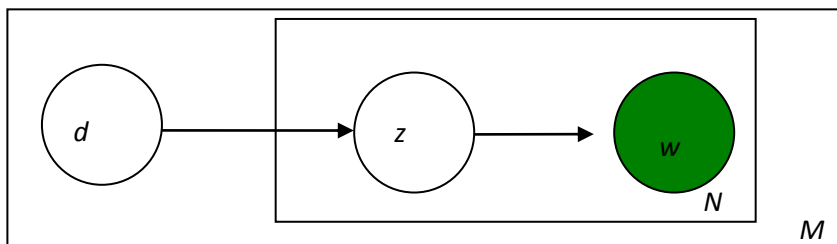


Figure 2.6. Plate notation of pLSA model.

In this plate notation, d is the document variable, z is a topic drawn from the topic distribution of this document, $p(z/d)$ and w is a word drawn from the word distribution for this topic $p(w/z)$. The observable variables are d and w and the latent variable is z . N is the number of words. M is the number of documents.

The objective of pLSA is to maximize the following log likelihood function.

$$L = \sum_{d \in D} \sum_{w \in W} n(d,w) \log(P(d,w)) \quad (4)$$

where $n(d,w)$ is the number of times word w appears in document d .

2.4.3 Relationship between LSA and pLSA

As previously mentioned, the statistical model for the pLSA is called the aspect model, a statistical mixture model. In this model, the observation pairs (d,w) are assumed to be generated independently and a “bag of words” approach is used in this model. In this model, a document d belongs to a set $D = \{d_1, d_2, \dots, d_n\}$, a topic z belongs to a set $Z = \{z_1, z_2, \dots, z_k\}$ and a word w belongs to a set $W = \{w_1, w_2, \dots, w_m\}$. A document is selected with a probability $p(d)$, a latent class z is selected with a probability $p(z/d)$, and a word is generated with a probability $p(w/z)$.

$$P(d,w) = p(d) p(w/d) \text{ where}$$

$$p(w/d) = \sum_z p(w/z) p(z/d) \quad (5)$$

The pLSA model is similar to LSA in that the probabilities expressed in the Aspect model can be expressed in matrix format. In LSA, a document term matrix is decomposed into three matrices by singular value decomposition.

$$\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^t \quad (6)$$

In pLSA, the probabilities in the aspect model can also be formulated in terms of matrices as follows.

$$\mathbf{U} =$$

	k=1	k=2	k=3	K=4
i=1	$p(d_1/z_1)$	$p(d_1/z_2)$	$p(d_1/z_3)$	$P(d_1/z_4)$
i=2	$p(d_2/z_1)$	$p(d_2/z_2)$	$p(d_2/z_3)$	$P(d_2/z_4)$
i=3	$p(d_3/z_1)$	$p(d_3/z_2)$	$p(d_3/z_3)$	$P(d_3/z_4)$

$$\mathbf{\Sigma} =$$

	k=1	k=2	k=3	K=4
k=1	$P(z_1)$	0	0	0
k=2	0	$P(z_2)$	0	0
k=3	0	0	$P(z_3)$	
k=4	0	0	0	$P(z_4)$

$$\mathbf{V}^t =$$

	j=1	j=2	j=3	j=4	j=5
k=1	$p(w_1/z_1)$	$p(w_2/z_1)$	$p(w_3/z_1)$	$P(w_4/z_1)$	$p(w_5/z_1)$
k=2	$p(w_1/z_2)$	$p(w_2/z_2)$	$p(w_3/z_2)$	$P(w_4/z_2)$	$p(w_5/z_2)$
k=3	$p(w_1/z_3)$	$p(w_2/z_3)$	$p(w_3/z_3)$	$P(w_4/z_3)$	$p(w_5/z_3)$
k=4	$p(w_1/z_4)$	$p(w_2/z_4)$	$p(w_3/z_4)$	$P(w_4/z_4)$	$p(w_5/z_4)$

$$\mathbf{X} =$$

	j=1	j=2	j=3	j=4	j=5
i=1	$p(d_1, w_1)$	$p(d_1, w_2)$	$p(d_1, w_3)$	$p(d_1, w_4)$	$p(d_1, w_5)$
i=2	$p(d_2, w_1)$	$p(d_2, w_2)$	$p(d_2, w_3)$	$p(d_2, w_4)$	$p(d_2, w_5)$
i=3	$p(d_3, w_1)$	$p(d_3, w_2)$	$p(d_3, w_3)$	$p(d_3, w_4)$	$p(d_3, w_5)$

Figure 2.7. Decomposition of pLSA into matrices.

Thus, \mathbf{U} is a matrix of $P(d/z)$ values, $\mathbf{\Sigma}$ = diagonal matrix of $P(z)$ values, and \mathbf{V}^t is a probability matrix of $P(w/z)$ values.

According to Lee et al. (2010), “words in a topic from pLSA are more closely related than words in a topic from LSA. For polysemy, words in a topic from pLSA can be appeared in other topics simultaneously.” pLSA partially handles polysemy. In pLSA, topics are multinomial random variables, each word is generated by a single topic, and different words may be generated from different topics. The limitation to pLSA is that there is no probability distribution model at the level of documents. Thus, the larger the number of documents, the larger the pLSA model.

2.4.4 The Latent Dirichlet Allocation Method

The latent dirichlet allocation method is a text data mining method that is an extension of the pLSA method. The main difference between the two methods is that key mixture probabilities now follow the dirichlet multinomial distribution which is given as:

$$Dir(\alpha_1, \dots, \alpha_T) = P(P_1, P_2, \dots, P_T / \alpha_1, \dots, \alpha_T) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma \alpha_j} \prod_{j=1}^T P_j^{\alpha_j - 1} \quad (7)$$

The plate notation for the LDA model is given by:

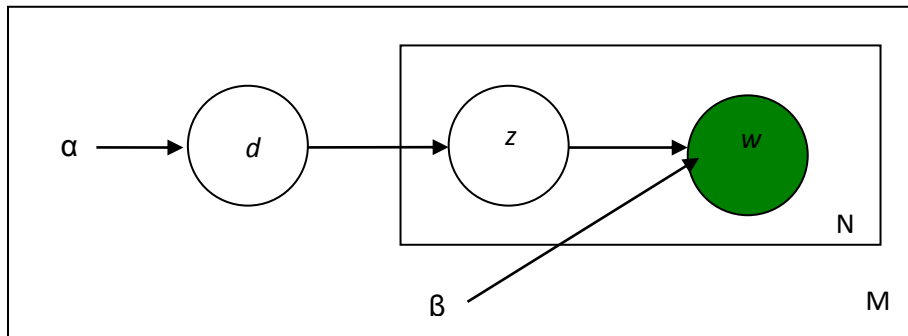


Figure 2.8. Plate notation on the LDA model.

As with pLSA, the latent dirichlet allocation method is based on the probabilistic topic model, the aspect model. Here, the probability of finding a word in a document is given by:

$$P(w_i) = \sum_{j=1}^T P(w_i/z_i = j)P(z_i = j) \quad (8)$$

where $P(Z_{i=j})$ =Probability that the topic= j for the i th word and $P(W_i/Z_{i=j})$ =conditional probability that of the word w_i given the fact that topic j was selected for the i th word. In terms of matrix notation, these probabilities are expressed as follows:

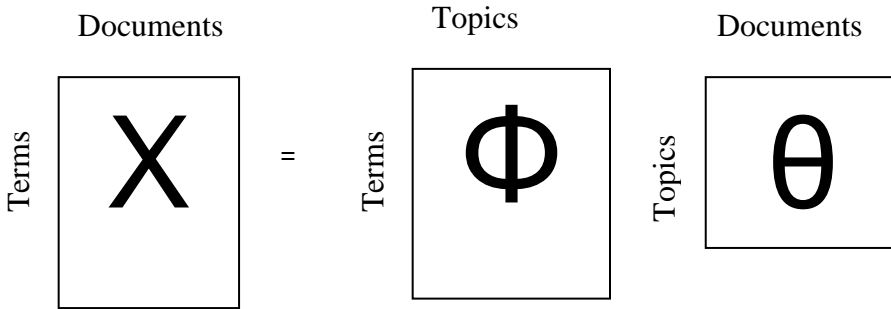


Figure 2.9. Decomposition of LDA into matrices.

Here, $\theta = P(z)$ and $\Phi = P(w/z)$. Blei, Ng, and Jordan (2003) defined $\phi^{(j)} = P(w / z=j)$ as the multinomial distribution over word for topic j and $\theta^{(d)} = P(z)$ refer to the multinomial distribution over topics for document d . Blei et al. (2003) and Griffiths and Steyvers (2004), extended Hofman's work by computing these probabilities through the sampling of specific mixtures that have probability distributions that follow the Dirichlet distribution where

$$Dir(\alpha_1, \dots, \alpha_T) = P(P_1, P_2, \dots, P_T / \alpha_1, \dots, \alpha_T) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma \alpha_j} \prod_{j=1}^T P_j^{\alpha_j - 1}. \quad (9)$$

With unknown parameters $\alpha_1, \dots, \alpha_T$, the formula for the Dirichlet Distribution is a formula that is very difficult to program in a computer. First, the range of every α_j parameter can exist

between $0 < \alpha_j < \infty$, yielding an infinite number of possibilities. The range for each probability value, P_i is from 0 to 1.0, with a constraint $\sum_{j=1}^T P_j = 1$. To make things easier, the Dirichlet is usually used with a constant parameter α or $\alpha_i = \alpha_j = \alpha$ for any i and j possibility. Using basic mathematics, Blei et al., (2003) decided to compute the posterior distribution of topics conditioned on given words as an alternative way to obtain the $p(w/\alpha, \beta)$. The rationale for this is that it is easier to start with words and determine topics rather than the other way around. By using several manipulations of Bayes' theorem, Blei et al. (2003) determined that the following equation for the posterior distribution of the topics z conditioned on the words was also intractable.

$$p(\theta, \mathbf{z} | w, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, w | \alpha, \beta)}{p(w | \alpha, \beta)} \quad (10)$$

Blei et al. (2003) also used the above conditional probability to determine the probability of selecting words $P(w/\alpha, \beta)$. Blei et al. (2003) assumed that both the prior $P(z)$ and the posterior probabilities $P(w/z)$ followed multinomial distribution from Dirichlet distributions mixtures with parameters (α, β) and concluded that this distribution $P(w/\alpha, \beta)$ that could have been used to generate documents is intractable and is given by the formula below:

$$p(w | \alpha, \beta) = \frac{\Gamma \sum_i \alpha_i}{\prod_i \Gamma \alpha_i} \int \left(\prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{i,j})^{w_n^j} \right) d\theta \quad (11)$$

Thus, because it is impossible to compute these probabilities directly when prior and posterior probabilities are derived from the Dirichlet distributions mixtures, different approaches have been used to estimate these key $p(W/\alpha, \beta)$ probabilities. The main issue in text mining has been how to determine the probability of finding a word in a document $p(W/\alpha, \beta)$. With LDA, the main issue is how to determine the probability of a word conditioned on the fact that $P(z)$ is a

multinomial probability from a *Dirichlet*(α) probability mixture and $P(w/z)$ is a multinomial distribution from a *Dirichlet*(β) probability mixture. Bayes' theorem and these probabilities are used to compute the probability of a word $P(W/\alpha, \beta)$.

2.4.5 Gibbs Sampling: Approximation to LDA

One of the approaches that are currently being used to solve the untractable probabilities of the latent dirichlet allocation method, is the Steyvers and Griffiths(2004) approach which consists of using a Markov chain Monte Carlo procedure called Gibbs sampling.

Hogg et al. (2005) explains the procedure for Gibbs sampling general procedure. To start out with, a stream of X_o values are initialized at time $t=t_o$. Then a conditioned random variable Y_i/X_{i-1} is generated from a distribution $f_{y/x}(y/x)$. This conditioned Y_i values are then substituted into $f_{x/y}(x/Y)$ distribution to generate a new set of conditioned X_i values or $X_i/Y_i \sim f_{x/y}(x/Y)$ and the process repeats itself. What is noted here is that the new state of the system only depends on the previous state and not on the past history. Also, the movement from one state to another is on a random basis. These two concepts are the basis for this Gibbs sampling procedure being called a Markov chain Monte Carlo—the future state only depends on the present state and not on its history and moving from one state to another state occurs on a random basis.

In this Gibbs sampling approach, Steyvers and Griffiths (2004) decided to define the posterior distribution of $P(z/w)$ as being proportional to the product of an estimate of $\varphi^{(j)} = P(w/z=j)$ and to an estimate of $\theta^{(d)} = P(z)$. In Griffiths and Steyvers's *Probabilistic Models* (2007) also defined the equation to select the next topic for a word token w_i given previous topics have been chosen for previous words as being proportional to these different terms and this equation is given below.

$$P(z_i = j | z_{-i}, w_i, d_i) \propto \frac{C_{w,j}^{WT} + \beta}{\sum_{w=1}^W C_{w,j}^{WT} + W\beta} \frac{C_{d,j}^{DT} + \alpha}{\sum_{d=1}^T C_{d,j}^{DT} + T\alpha} \quad (12)$$

The first term represents an estimate of $\varphi^{(j)} = P(w/z=j)$ and the second term represents an estimate of $\theta^{(d)} = P(z)$. They suggested that a good value for $\alpha = 50/T$, where T =number of topics and $\beta=0.01$. In this formula, C^{WT} and C^{DT} are matrixes with dimensions W by T and D by T respectively. C^{WT}_{wj} contains the number of times word w is assigned to topic j , not including the current instant i and C^{DT}_{dj} contains the number of times topic j is assigned to some work token in document d , not including the current instance i . The distributions $P(z)$ and $P(w/z=j)$ can be obtained from the C^{WT} and C^{DT} matrixes, either after each C^{WT} and C^{DT} is being computed or after a specific number of iterations have updated the C^{WT} and C^{DT} matrixes.

The estimate probabilities $\varphi^{(j)} = P(w/z=j)$ and $\theta^{(d)} = P(z)$ used to compute the probability of finding a word w_i are given by:

$$\Phi_i^{(j)} = \frac{C_{w,j}^{WT} + \beta}{\sum_{w=1}^W C_{w,j}^{WT} + W\beta} \quad (13)$$

$$\theta_j^{(d)} = \frac{C_{d,j}^{DT} + \alpha}{\sum_{d=1}^T C_{d,j}^{DT} + T\alpha} \quad (14)$$

How the above equations follow the Gibbs sampling procedure is briefly outlined here. At time $t=t_o$, the topics are selected randomly for a word w_i in a document d_j . From these initial values the C^{WT} and C^{DT} matrixes are computed. These matrixes are then updated to compute the terms. $\varphi^{(j)} = P(w/z=j)$ and $\theta^{(d)} = P(z)$. Then these terms are used to compute the probability of selecting the next topic for w_i in a document d_j or $P(z_{i=j}/z_{-i}, w_i, d_i, \dots)$. Once the topics have been updated for each word w_i , new matrixes C^{WT} and C^{DT} are computed and the process keeps going

until it converges and there is not much difference in the matrices after a number of iterations. Appendix I lists the flowchart of how this algorithm works.

2.5. Conclusion

The main text mining techniques that will be covered in this dissertation are the LSA and LDA. Because LDA is commonly recognized as an improvement over pLSA and pLSA is considered to be an outdated technique, pLSA is not covered. For the research covered in this dissertation, algorithms have been developed to implement the LSA and LDA techniques. The algorithm for LDA is based on Steyvers and Griffith's implementation of the Markov chain Monte Carlo Gibbs sampling theory and this algorithm is considered a good technique to obtaining LDA results.

CHAPTER 3

CLASSIFICATION

3.1 Classification as Supervised Learning

Classification is a form of supervised learning. The process of assigning pre-defined category labels to a new database tuple based on the analysis of a labeled training data set is called classification (Han and Kamber, 2006). In contrast, the process of assigning a continuous or ordered value to the database tuple from a labeled training data set is called prediction.

Classification and prediction are used to analyze large databases in many settings. They are used for fraud detection, target marketing, performance prediction, medical diagnoses, etc.

Supervised learning is different from unsupervised learning and the objectives under each different mode of data analysis are different. The goal of unsupervised learning is to identify clusters (or groups) in an unstructured data set and to discover the structure that may exist in the data set. The process involves applying several techniques to identify groups of similar objects that may exist in the data set. In supervised learning which consists of classification and prediction, the situation is different. The data set has labeled data points where each label, ω_k , belongs to a set of c classes or $\omega_k \in \{1, 2, \dots, c\}$ where c is the number of classes. Again, if the classifier, Φ , predicts categories (discrete, unordered labels), then the process is called Classification. If the classifier predicts continuous numeric values, then the process is called Prediction.

Reinforcement learning is a special case of supervised learning where an additional label, called a reinforcement signal $r(w)$, is added to the data set that can be used as a binary signal.

For example, if c classes are available in the data set, a reinforcement signal $r(w)$ can be added to the data set such that

$$r(w) = \{ 1 \text{ if class label is even, and } -1 \text{ if class label is odd } \}$$

Thus, supervised and unsupervised learning are two extremes and reinforcement learning falls somewhere in between these two modes of learning.

3.2 Data Points Representation

Usually data points are represented by vectors in a vector space and the supervised and unsupervised processes involves identifying the data points by either measuring the distance between the points or the similarity between the data points. There are many ways of measuring the distance between two data points. One frequently used category of distances is known as Minkowski distance category. Cios et al. (2007), list the following distance metrics as part of this category.

$$\text{Hamming distance} \quad d(x,y) = \sum |x_i - y_i| \quad (15)$$

$$\text{Euclidean distance} \quad d(x,y) = \sqrt{(\sum (x_i - y_i)^2)} \quad (16)$$

$$\text{Tchebyshev distance} \quad d(x,y) = \max (|x_i - y_i|) \quad (17)$$

Another important mathematical concept that is used to determine the similarity between two vectors is the dot product. The dot product of two vectors **A** and **B** is given by the following formula:

$$A \cdot B = |A| |B| \cos (\theta) \quad (18)$$

This formula is important for document retrieval because in text mining, documents are usually expressed in what is known as the vector space model (VSM), as vectors of terms (Salton 1975).

The importance of the term in each document is determined by its frequency of appearance in the document.

3.3 What Are Classifiers

What are classifiers? There is not a consistent definition for a classifier. Classifiers can be algorithms, decision rules, mathematical formulas, and any other technique that can discriminate between classes of patterns. There are two-class and multi-class classifiers. The type of classifier used depends on number of classes, the type of data, the learning algorithm used to develop the classifier and the procedures to validate the classifier.

The mapping of classifier Φ is denoted as:

$$\Phi: X \rightarrow \{ \omega_1, \omega_2, \dots, \omega_c \} \quad (19)$$

where the database is X whose elements are \mathbf{x} and the classes available are $\omega_1, \omega_2, \dots, \omega_c$.

What are possible classifiers? There are linear and nonlinear classifiers. Classifiers can be decision tree classifiers, Bayesian classifiers, Bayesian belief networks, support vector machines (SVM), k-nearest neighbor (k-NN), centroid-based, case based reasoning, genetic algorithms, rough sets, fuzzy logic techniques, etc. Methods for prediction include linear and nonlinear regression. In many of these classifiers, the user's experience and knowledge determines the final parameters of the classifier used. For example, to use the k-NN classifier, the user has to specify which "k" value to use and which distance metric to use.

3.4 Developing a Classifier

The development of a classifier usually starts with the existence of a raw data set that will eventually be analyzed by the classifier. Before a classifier is developed, this data set is usually cleaned. A data set consists of labeled and/or unlabeled data tuples, otherwise known as

examples, instances, or data points. A labeled data tuple allows for supervised learning data analysis. An unlabeled data tuple is designed for unsupervised learning data analysis. The data set is cleaned to remove “noise” and to treat missing values in the data set. In attempting to clean the data, the correlation that may exist between attributes in the data also needs to be determined. If two attributes in the data set are highly related, one attribute may have to be deleted from the data set as it may not contribute additional information to the data set or to the classification/prediction procedure. Lastly, the data also needs to be analyzed for possible data transformation (e.g., Should numeric values for income be transformed to “medium”, “low”, or “high” income?).

A database D has m observations in terms of $n-1$ explanatory attributes (categorical and/or numerical) and a categorical target attribute. In this database, a tuple X is represented by an n -dimensional attribute vector $X = (x_1, x_2, x_3, \dots, x_n)$ where x_1, x_2, \dots, x_n are n measurements made on the tuple from n -database attributes A_1, A_2, \dots, A_n . Each tuple X is presumed to belong to a discrete unordered class label, another database attribute. The database consists of many tuples that can be divided into three different data sets: the training set, the validation set, and the testing set. The division of the data is usually done after the data has been “cleaned” to the satisfaction of the analyst. The learning step of building a classifier is usually achieved through an algorithm that builds the classifier by “learning from” the training data set. According to Vercellis et al. (2009), “from a theoretical viewpoint, the development of algorithms capable of learning from past experience represents a fundamental step in emulating the inductive capabilities of the human brain.” The validation set is usually used to measure the classification accuracy of the classifier created from the training data set and to detect if the classifier has been made too complex and/or impractical. For example, it is through the validation data set that a

higher order polynomial type of classifier can be simplified by adjusting the order of the polynomial. Thus, the validation data set serves to modify the original classifier to make it more effective for classification purposes. Finally, the testing set is used to measure the effectiveness of the final chosen classifier. The performance of the classifier is measured by its accuracy, where accuracy is usually the percentage of the test data tuples that are correctly classified by the classifier. As mentioned before, the difference between a classifier and a predictor is that the output y is usually a continuous and ordered value for a predictor and a labeled value for a classifier.

3.4.1 Loss Functions

The development of a classifier is usually done through the minimization of a loss function. Vapnik (2005) defined the objective function of a Support Vector Machine classifier as being the expected risk on a test sample as being the training-set error (misclassification rate) and the inverse of the margin width. The margin width measures how far two different types of examples of a category are separated by the decision surface. The following figure serves to illustrate how one classifier may differ from another based on the expected risk.

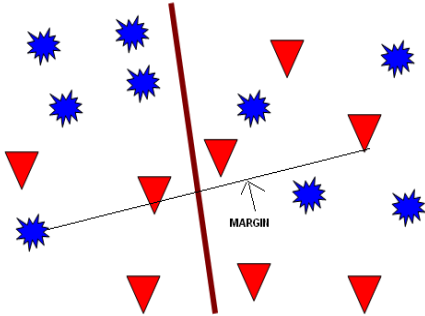


Figure 3.1: Classifier with 41% misclassification rate.

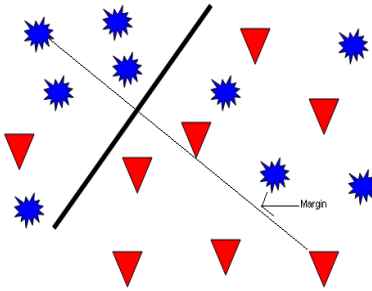


Figure 3.2: Classifier with 29% misclassification rate.

To be able to compare different classifiers on the same basis, Li and Yang (2003) tested and defined a loss function for nine classifiers. This loss function consisted of having a training-set

loss term and the regularizer (complexity penalty) term. The following table is adapted from this article:

Table 3.1:

Training Loss Functions and Regularizers for Eight Classifiers (from Li and Yang, 2003)

Classifier	Training-set loss: $g_1(y_i \bar{x}_i \bar{\beta})$	Regularizer: $g_2(\bar{\beta})$
Regularized LLSF	$\sum_{i=1}^n (1 - y_i \bar{x}_i \bar{\beta})^2$	$\lambda \ \bar{\beta}\ ^2$
Regularized LR	$\sum_{i=1}^n \log(1 + \exp(-y_i \bar{x}_i \bar{\beta}))$	$\lambda \ \bar{\beta}\ ^2$
Regularized 2-Layer NN	$\sum_{i=1}^n (1 - \pi(y_i \bar{x}_i \bar{\beta}))^2$	$\lambda \ \bar{\beta}\ ^2$
SVM	$\sum_{i=1}^n (1 - y_i \bar{x}_i \bar{\beta})$	$\lambda \ \bar{\beta}\ ^2$
Rochio	$\sum_{y_i=1}^n (y_i \bar{x}_i \bar{\beta}) - \frac{b N_c}{N_c} \sum_{y_i=-1}^n y_i \bar{x}_i \bar{\beta}$	$\frac{N_c}{2} \ \bar{\beta}\ ^2$
Prototype	$\sum_{y_i=1}^n (y_i \bar{x}_i \bar{\beta})$	$\frac{N_c}{2} \ \bar{\beta}\ ^2$
kNN	$\sum_{y_i=1 \wedge \bar{x}_i \in R_k(\bar{x})}^n (y_i \bar{x}_i \bar{\beta})$	$\frac{1}{2} \ \bar{\beta}\ ^2$
NB without smoothing	$-\sum_{y_i=1}^n (y_i \bar{x}_i \bar{\beta})$	$S_c \ \bar{\beta}\ _1$
NB with Laplace smoothing	$-\sum_{y_i=1}^n (y_i \bar{x}_i \bar{\beta})$	$(p + S_c) \ \bar{\beta}\ _1 + \ \bar{\beta}\ _1 +$

In this table y_i =labeled data point, β consists of parameters of the linear classifier, x_i is input values of the training data set (a tuple), N is the number of data points $(x_1, y_1) \dots (x_N, y_N)$, N_c is the number of data points in a category, and λ is a constant.

3.4.2 Data Selection

The selection of the data chosen to develop the classifier can make a significant different in the classifier's effectiveness. According to Cios et al. (2007), "a classifier that is being trained and evaluated on one data set produces overly optimistic results that could lead to catastrophic behavior of the classifier in practice." To achieve a certain level of confidence on the classifier, then a classifier needs to be created from an appropriate data set. Sometimes excellent performance of the classifier on the training set can often result in poor relative performance in

the testing set. This effect is called “memorization.” Thus, classifiers do not always generalize well, especially if they are designed to be too complex. Sometimes, they can be effective in the data that they were trained on, but they cannot be used to generalize beyond this data set. One possible reason for classifiers behaving this way is that noisy data may have been incorporated into the original training set.

3.5 Classifier Performance

The performance of a classifier is a function of the type of data, the background and experience of the user, the implementation method and the use to which the classification method will be put (Jamain et al., 2008). In their classification meta-analysis, Jamain et al. (2008) explain that an expert who is familiar with the parameters of a classification method has a better “feel” for the classification rules and thus is more likely to get better classification results than a novice user.

3.5.1 Classification Matrix

The performance of a classifier can be measured in many ways. One of these ways is through a classification matrix. Shmueli et al. (2007) defined the classification matrix for a binary classifier in the following manner:

Table 3.2

Classification Matrix

	Predicted Class	
	Co	C1
Actual Class Co	$n_{0,0}$	$n_{0,1}$
Actual Class C1	$n_{1,0}$	$n_{1,1}$

- $n_{0,0}$ = Number of Co cases correctly classified as Co
- $n_{0,1}$ = Number of Co cases incorrectly classified as C1
- $n_{1,0}$ = Number of C1 cases incorrectly classified as Co
- $n_{1,1}$ = Number of C1 cases correctly classified as C1

Here, the accuracy of the classifier is measured by:

$$\text{Accuracy} = \frac{n_{0,0} + n_{1,1}}{n_{0,0} + n_{0,1} + n_{1,0} + n_{1,1}} \quad (20)$$

Many algorithms compute for an example the probability of belonging to a particular class. For these types of situations, a cutoff value is deemed to be the classifier. For example, if the probability \geq cutoff value, then a test example basically belongs to class “A”, else it belongs to class “B”. The chosen cutoff value determines the level accuracy and misclassification rates.

Shmueli, et al. (2007), also defined other parameters in terms of the variables used in the classification matrix. The sensitivity of a classifier detects the ability of the classifier to detect the important class member correctly. Sensitivity is defined as:

Sensitivity = $n_{0,0}/(n_{0,0} + n_{0,1})$ = Fraction of all class C₀ members that are correctly classified.

Specificity detects the ability of the classifier to detect another particular class correctly.

Specificity here is defined as:

Specificity = $n_{1,1}/(n_{1,0} + n_{1,1})$ = Fraction of all class C₁ members that are correctly classified.

Other definitions of interest are:

The fraction of falsely predicted positive out of all predicted positives is

$$\text{False positive rate} = (n_{1,0})/(n_{0,0} + n_{1,0}) \quad (21)$$

The fraction of false predicted negative out of all predicted negatives is

$$\text{False negative rate} = (n_{0,1})/(n_{0,1} + n_{1,1}) \quad (22)$$

3.5.2 Receiver Operating Characteristic (ROC) Curves

Another measure of performance for a classifier is the receiver operating characteristic (ROC) curve which plots the sensitivity versus the specificity. The area under the curve (AUC) is a measure of the performance of a classifier. The bigger the AUC value, the better the classifier. The (sensitivity, specificity) values can change as function of the cutoff values of a

classifier. They can also change as function of parameters of a classifier. Also, these data points can change as a function of the size of the database that is being classified. The plotting of ROC curves for different classifiers allow for the comparison of classifiers for a range of database sizes.

3.5.3 F_1 Micro-Sign and F_1 Macro Sign Tests

Another measure to evaluate the performance of a classifier are the F_1 -micro-sign test and the F_1 -macro-sign tests. Yang and Liu (1999) used these two statistics to compare two systems based on binary decisions on all test documents/categories. According to Tam et al. (2002), in these tests, precision, p , is defined as the percentage of retrieved documents that are relevant, and recall, r , is defined as the percentage of relevant documents retrieved. The formula for the F_1 statistic for both tests is:

$$F1(r, p) = 2rp/(r + p) \quad (23)$$

When F_1 is computed over each individual category and then averaged over all categories, the resulting F_1 statistic is used for macro-averaging. When F_1 is computed globally over all tests documents, the computed statistic is F_1 micro-sign.

Gopal and Yang (2010) also listed the following F_1 metric for a multi-category situation where a classification matrix is created for each category that is tested. In this situation, the results of each category are divided into four groups: True Positives (TPc), False Positives (FPC), True Negatives (TNc), and False Negatives (FNc). The corresponding metrics for this situation are

$$\text{Global Precision P} = (\sum_{c \in C} TPc) / \sum_{c \in C} (TPc + FPC) \quad (24)$$

$$\text{Global Recall R} = (\sum_{c \in C} TPc) / \sum_{c \in C} (TPc + FNc) \quad (25)$$

$$\text{Micro-average } F_1 = 2 * R * \frac{P}{P + R} \quad (26)$$

To compute the Macro-average F_1 , the following parameters are computed for each category $c \in C$

$$C = \{c_1, \dots, c_m\}.$$

$$\text{Category-specific Precision} = P_c = TP_c / (TP_c + FP_c) \quad (27)$$

$$\text{Category-specific Recall} = R_c = TP_c / (TP_c + FN_c) \quad (28)$$

$$\text{Macro-Average } F_1 = \{ \sum_{c \in C} (2P_c R_c / (P_c + R_c)) \} / m \quad (29)$$

3.5.4 Misclassification Costs

There is usually a cost (or benefit) associated with classification. But one thing to notice is that misclassification also carries subjective severity implications that vary per application. Misclassification rates for hospital treatments can yield to personal injury and even deaths of patients but for college admission, misclassification may just cause a loss of admission. The following is an illustration of how the cost (or benefit) associated with misclassification can be computed. Suppose that there is a situation where a profit associated with classifying a user as class 1 is \$10 and the cost of sending material to a user of \$2.00. Also, suppose that there exists a policy of only sending material to users who are predicted to belong to class I. For the following dataset of 50 users being classified as follows by a specific classifier:

Table 3.3

Sample Classification Matrix

	Predicted Class	Predicted Class
	Co	C1
Actual Class Co	10	6
Actual Class C1	20	14

The cost matrix is given by:

Table 3.4

Cost Classification Matrix

	Predicted Class Co	Predicted Class C1
Actual Class Co	\$100	-60
Actual Class C1	-40	0

Average cost of misclassification=cost of misclassifying class I * number of class I misclassifications + cost of misclassifying class II * number of class II misclassifications.

In the above example, the cost of misclassification is = -40 + -60 = \$-100.00 . Or the expected loss was \$100 dollars for misclassification.

3.6 Document Classification

Why is it desired to classify documents? According to Tam et al. (2002), classification can facilitate document retrieval. Klimt et al. (2004) explained that if the documents are e-mails, the documents are classified to determine whether an e-mail is spam, to determine which folder to assign an e-mail to, and to determine the priority of e-mails. But the classification of documents is not problem free. In document classification, the attributes are words or tokens and the labels are topics. In determining a classifier, “learning” occurs from one data set and applied to another data set. The assumption is that the new data set and associated examples are supposed to be *i.i.d.*, (independently and identically drawn) as the first data set. However, in many data sets, this assumption does not hold. In particular, if the data set consists of news stories, the data set is expected to be dynamic. The type of stories that will be collected over time will change over time and one data set is not expected to be equivalent to another data set.

But categorization can still be applied and the assumption is that the nature of the data set is supposed to remain relatively stable over a period of time.

3.7 Commonly Used Document Classifiers

There are many document classifiers and it is difficult to cover and list them all in detail. However, three classifiers that seem that stand out in the literature is the k -nearest neighborhood (k -NN) classifier, the support vector machine (SVM) classifier and the centroid-based classifier. There have also been many studies to compare the methods. Tam et al. (2002) compared the performance of the k -nearest neighborhood (k -NN) algorithm, the centroid-based classifier, and the highest average similarity over retrieved documents (HASRD) algorithm as document classifiers. The measure of performance used to compare these algorithms was the F_1 micro-sign test and the F_1 macro-sign test. The results indicate that k -NN performed the best, followed by an adapted HASRD, and last was the centroid-based classifier.

How a test document is assigned to a category depends on the algorithm used. In the HASRD algorithm, the documents with the highest similarity (using the dot product formula) to the test document are retrieved. Then the average similarity of M returned documents for each category is computed. This average similarity value is regarded as the similarity between the test document and category c_j . The process is repeated for several categories. The test document is assigned to the category that exhibits the largest average similarity value.

3.7.1 k -Nearest Neighborhood Classifier

In the k -nearest neighborhood (k -NN) algorithm, a test document, d , is compared to documents in the training set that belong to a certain category. In each category, the algorithm compares the k nearest neighbor documents to the test document at hand and computes the similarity value, (based on the dot product formula). The average similarity value for that

category can be computed and/or the cumulative similarity value can also be used to determine to what category the test document, d , needs to be assigned to. In determining which are the nearest neighbors to a test data point, the user needs to determine which metric distance to use and also how many data points, k , to consider to be classified as neighbors of the test document.

3.7.2 Centroid-Based Classifier:

In the centroid-based algorithm, the training documents are grouped together based on category c_j . The centroid of each category group is computed using the formula:

$$C_j = (1/|D_j|) \sum_{d \in D} d \quad (30)$$

where C_j is the vector representation for category C_j , D_j = number of documents in the training set which belong to category C_j , and d is the vector of weights representing document, d , in the training set. In this algorithm, given a test document, d , the goal is to compare the similarity between d and the C_j value for each category. The dot product formula can be used to compare d and the C_j value. The category with the highest similarity value will be the category assigned to document d .

3.7.3 Support Vector Machine

Another popular document classification technique is the support vector machine (SVM) classifier. The SVM is an algorithm that creates a hyperplane as a decision vector and the basis for selecting this hyperplane is the one that has the wider margin that is supported by data points called support vectors. The data points that touch the margin are called the support vectors.

Figure 3.3 illustrates the main idea behind Support Vector Machine classifiers.

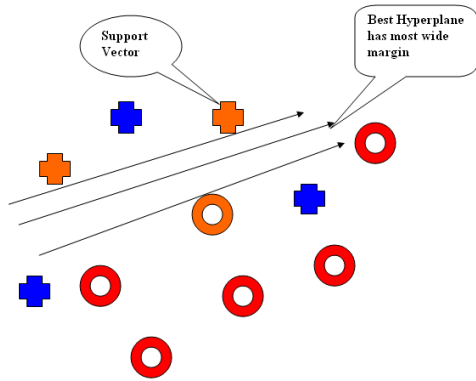


Figure 3.3. Support vector machine illustration.

Klimt, et al. (2004), used the support vector machine as a classifier of e-mails from an Enrod corpus. What was unique in this research is the way that they created the training data set to develop the classifier. The authors basically split an e-mail message into two parts. The top part was used for the training set, and the bottom part was used for the testing set. The evaluation of the classifier was used by computing the F_1 scores (micro-average, and macro-average) on the number of messages by a user, and then on the number of folders used by the user. In this research they discovered that the body of the e-mail message was the part of the e-mail that was most useful as a classifier. They also discovered that the threading of e-mails can present a problem for text mining. Threading is the process of one e-mail being related to previous e-mails. The problem with threading is that later e-mails do not always cover the original topics that initiated the original e-mail.

3.8 Conclusion:

Many different classifiers exist to evaluate corpus of documents. They each vary in terms of format and effectiveness. In the proposed research for this dissertation, two text mining techniques, LDA and LSA, were used as classifiers of documents. LDA and LSA algorithms have already been developed and as a result, they do not need to go through the regular

development process of a new classifier (e.g., start with a training data set, etc.). However, LDA and LSA are not fixed algorithms. For LSA, the number of dimensions to use in the final truncated matrix is a subjective choice that depends heavily on the experience of the user. In LDA, the α and β parameters of the Dirichlet distribution are also adjustable parameters. In this classification research, it is expected that classification matrices will vary for each type of adjustment made in each of these algorithms and this is expected to affect the overall classification performance of the algorithms.

CHAPTER 4

HUMAN CATEGORIZATION LITERATURE REVIEW

4.1 Introduction

The ability for humans to categorize is considered to be a foundation of cognition in human categorization literature (Arterberry and Bornstein, 2002). The field of human categorization is vast and it has been studied in the literature from many different aspects. Human categorization studies range from childhood cognition developmental studies to simulation modeling of the mechanisms of human categorization mental processes to understand the intricacies involved in the human brain. In many human developmental studies, the consensus is that humans respond to stimuli presented to them from the beginning of their childhood. Usually, this stimulus is visual, but it can also be sensory, and auditory. As humans mature, they learn to distinguish entities and group similar items through features that they find to be linked in a common manner. Infants at first learn to categorize simple forms, then animals and nonliving objects. Eventually, they recognize gender of faces and emotional expressions. They learn to categorize faces after recognizing specific facial features, (e.g. “mustache,” “eyes,” “nose,” etc).

The field of human categorization has been divided in two main areas of study: category learning and category representation. According to Ashby and Maddox (2005), the brain’s neural mechanism that mediates the learning of different categories is different from the brain’s neural structures that mediate the representation of learned categories. As evidence of the existence of this difference, Ashby and Maddox (2005) mention that frontal damaged patients and Parkinson’s disease patients that are impaired in category learning can still perform category

representation tasks by recognizing old familiar categories (e.g., animals, fruits, etc.) and no evidence exists that impairment in category representation (or failing to recognizing old familiar categories or development of agnosias) can lead to general learning category problems.

Regarding categorization representation, Bornstein (1984) proposed that four types of categorization representations occur in the evolution of man: Identity categorization, reference equivalence categorization, perceptual equivalence categorization and conceptual equivalence categorization. Identity categorization refers to the recognition of the stimulus across multiple representations (e.g., Grandma would always be recognized as Grandma even if she were to dress up in a funny Halloween suit). Reference equivalence categorization is recognizing the stimulus across variations of appearances and from different angles of appearances (e.g., a child would recognize his father even if he/she is looking at the back of his father's head). Perceptual equivalence categorization is grouping different and discriminative stimuli by their qualitative attributes (e.g., trucks, cars, and buses are grouped into vehicles because they have wheels). Conceptual equivalence categorization considers the common dimensions in different stimuli and uses them to group them together items (e.g., a cow, a dog, and a man are grouped together into the category mammals). According to Arterberry and Bornstein (2002), infants tend to categorize at the perceptual and conceptual level. Perceptual categorization relies heavily on the physical appearance of a stimulus and conceptual categorization relies on the functions and roles that stimuli play in events.

4.2 Categorization at the Developmental Stages

In trying to determine how categorization learning evolves in infants, interesting research has been conducted. Arterberry and Borstein (2002) conducted research on infants, ranging from 6 to 9 months old, and tried to determine if infants trained to categorize static objects could

transfer their categorization skills to dynamic objects and vice versa. Experiments were conducted where the training stage was followed by a testing stage. In the first case, static to static, the training stage consisted of infants looking at static images on a computer screen followed by a testing phase where infants were presented again with static images of novel items as well as habituated items from the training stage. The infants' ability to categorize was determined by measuring the length of time that the infants spent looking at an object on a projected computer screen. The conclusion was that 6 month old infants preferred new items over the habituated items or they spent more time looking at new items over habituated items in the testing phase of the static training to static testing case (Arterberry and Bostein, 2002). Similar outcome was obtained when infants were tested in the dynamic training to dynamic testing case. Here, the dynamic stage consisted of infants looking at dynamic point light displays that showed similar images of the static pictures. Arterberry and Borstein (2002) then conducted the static training to dynamic training case and the vice versa case. For these two cases, 6 month old infants did not show any significant preference for new items in the testing phase. This indicated that to the 6 month old infants, all items (old habituated ones and the new ones) displayed in the second test stage seemed "new" to them. They simply could not make any distinction between the old habituated items and the new items in the testing stage. However, when 9 month olds were tested, Arterberry and Borstein (2002) discovered that these older infants could make a distinction when tested under the dynamic training and the static testing case, but not vice versa. The results showed that 9 month old infants habituated to static cues, showed no ability to transfer their categorization skills to dynamic cues. But if they are habituated to dynamic cues, they showed that they could transfer their categorization skills to static cues. Arterberry and Borstein (2002) concluded that the 9 month old infants' ability to

transfer category information from the dynamic to the static stage show that they have developed more their conceptual abilities to categorize and were not just relying on their perceptual abilities.

4.3 Category Learning Theories

Three main category learning theories predominate in the human categorization literature: decision bound theory, prototype theory and exemplar theory. Decision bound theory assumes that categories belong to areas in a multi-dimensional stimulus psychological space. The features associated with a category (e.g., size, weight, etc.,) are used as dimensions in the stimulus psychological space. Decision bound partition the stimulus space into response regions by establishing boundaries. These boundaries can be linear and nonlinear created by the combinations of many different dimensions of categories. When an unfamiliar stimulus is encountered, the subject determines under what region it falls under and emits the associated response. Category learning under decision bound theory is the process of learning the regions enclosed within the boundaries (Ashby & Gott, 1988; Ashby & Townsend, 1986). Prototype theory is another category learning theory that explains that categorization is made in reference to a central category member, a prototype. A category can have different exemplars, versions, of the central prototype. When an unfamiliar stimulus is encountered, it is assigned to the category with the most similar prototype (Homa et al., 1981; Posner & Keele, 1968, 1970; Smith & Minda, 1998). According to Ashby and Maddox (2005), "Prototype theory assumes a category is represented as a prototype, and stimuli are categorized by comparing them to the prototypes of each contrasting category." Exemplar theory is another category learning theory that also assumes that there are many versions (exemplars) that exist within category. When an unfamiliar stimulus is encountered, it is compared to *every* exemplar that exists. The stimulus is

assigned to the category where its similarity is the highest in terms of the number of exemplars within a category (Brooks, 1978; Estes, 1986, 1994; Hintzman, 1986; Lamberts, 2000; Medin & Schaffer, 1978; Nosofsky, 1986). To better illustrate the difference between these two latter learning theories, Ashy and Maddox (2005) provided an analogy of how prototype theory and exemplar theory differ by using the solar system, in which the sun is the central prototype and the planets are distortions (exemplars) of the central prototype. The probability of an individual responding “A” in category task of the type “(A, not A)” can change with the position of the stimulus in the solar system space model. Under prototype theory, the probability of responding “A” depends on the distance between the stimulus and the central prototype, the sun. With exemplar theory, the probability of responding “A” depends on the sum of the distance between the stimulus and all of the exemplars, the planets.

4.4 Human Categorization Modeling

To understand the mechanism for human categorization learning, different simulation models have been created that have evolved over time. Appendix A gives a detailed view of the functionality of some of the models that have been covered extensively in the field of human categorization. One of the first models that have been described in the literature is the basic-cue categorization model that was developed by Gluck and Bower (1988). This model is a network model where input nodes are provided for every feature of a stimulus and output nodes represent the categories available. Each input node is linked to an output node through a weighed connection. The input nodes are represented as points in a multidimensional psychological space with specific values assigned to each dimension. Kruschke (1994) enhanced the basic-cue categorization model by introducing a new shifting attention mechanism to the input stimuli nodes and created the attention to distinctive input (ADIT) model. Kruschke (1992) also

developed the attention learning covering map (ALCOVE) model which differed from the ADIT model in that it consisted of a three layer hierarchical learning model that relates a stimulus to an “exemplar” (denoted as hidden node in the network) and then to the output category node. The idea behind this model is that categories can exist in different versions, exemplars, with different probabilities of occurrences and participants make categorization decisions based on the manner in which they relate to exemplars. This model was developed in an attempt to explain the exemplar theory of categorization and exemplar nodes are activated based on the psychological distance between the exemplar and the input stimulus node. This model incorporates the Nosofsky (1986) similarity function from the generalized context model (GMC) to measure this psychological distance.

Simulation studies of human categorization learning have also incorporated casual theory. Rehder (2003) recognized that certain features can be linked through casual relationships. Rehder (2003) gives an example of casual model theory where he explains that birds have wings, can fly, and can build nests up in trees but these features are linked in the following manner. Birds can fly *because* they have wings. Birds can build nests up in the tree *because* they can fly. Rehder (2003) research discovered that there are certain features that are more salient and affect categorization and that features that are linked through a casual relationship affect categorization accuracy more than features that are not linked.

4.5 Categorization Issue: Base Rates (Prior Probabilities)

In the study of human categorization learning, base rates (prior probabilities) of categories are factors that are believed to affect the categorization process. Kruschke, J. K. (1996), explains that humans utilize knowledge of category base rates to make category judgements. According to Kruschke (1996), base rates have two roles in the study of

categorizations. The primary role of base rates is to cause the high frequency categories to be learned before the less frequent categories. The second role for base rates is to bias the subjects to choose the most frequent categories. In his study, Kruschke (1996) confirmed in an experiment with different base rates that the most frequent categories are learned first before the rare categories, that subjects were able to recognize that categories occurred with different base rates during the experiment, and that subjects generally showed bias toward the most common categories. In another experiments with categories having the same rate, Kruschke (1996) discovered that participants chose the category that the participants learned last.

4.6 Categorization Issue: Inverse Base Rate Effect (IBRE)

But in the study of human categorization, one phenomenon that has arisen and is the subject of numerous studies is the inverse base rate effect (IBRE). IBRE is the name given to the observation of participants choosing the rare category over a common category when confronted with conflicting cues and they choose the rare category in a proportion that is related to their inverse rate of occurrence. An example of how the inverse base rate effect is studied in the literature follows. Three symptoms (dizziness, headaches, and skin rash) are used to predict two diseases: the common disease C and the rare disease R. If during training, the participants learn that dizziness (I) and headaches (PC) are symptoms of common disease C, and dizziness (I) and skin rash (PR) are symptoms of rare disease R, the participants can easily use the combination of dizziness and headaches, (I, PC) to predict C, the common disease, and the combination of dizziness and skin rash (I, PR) to predict R, the rare disease. In making these predictions, participants shift their attention to the symptoms headaches (PC) and skin rash (PR) (rather than dizziness (I)) to make these predictions. (This simple attention shift clearly indicates that humans have an attention shifting mechanism that allows them to recognize that not all cues

carry the same attention weight when making predictions.) However, when participants are given the *conflicting* combination of headaches and skin rash (PC, PR), most participants predict the rare disease, R, contrary to the frequency of occurrence for the rare disease and in a proportion that is related to their inverse rate of occurrence. To illustrate how the inverse rate effect works, the following scenario is given. If common disease *C* is present 75% of the time in the population and rare disease *R* is present 25% of the time, the ratio of base rates is 3 to 1. The inverse base rate effect basically states that participants will choose category *C* 1/3 of the time and category *R* 2/3 of the time. Kruschke (1996) experimental research also confirmed the IBRE effect in an experiment where the common disease/rare disease ratio was 3:1. Here, participants when confronted with two conflicting cues (PC and PR), participants chose the rare disease with a proportion 61.2% of the time. Kruschke(1996) repeated the experiment with another set of rare and common disease and discovered that when participants were confronted with conflicting cues, they chose the rare disease 56.3% of the time.

Why do people tend to predict the rare category more than the common category in an inverse proportional level when confronted with conflicting cues? Researchers attempt to find explanations for this inverse base rate phenomenon. One explanation is that the people's knowledge of asymmetry of the categories causes this IBRE phenomenon (Krushne, 2001). Asymmetry of the categories is that there is a higher probability for one relationship to exist over another one. Another explanation is called the eliminative inference approach which assumes that people form inference rules that are used in a flexible and controlled manner to eliminate category membership. Medin and Edelson (1988) used this approach to reproduce the basic rate effects found in an inverse base rate design. Another explanation that exists is grounded on Attentional Theory and claims that rapid attention shifts that occur during the learning phase of

categorization causes IBRE. It is believed that if participants, during the training of category selection, they attempt to learn the categories as fast as possible and with as few trials as possible, they are more likely to choose the rare categories when confronted with conflicting stimuli (e.g., PC, PR).

As to what is the best explanation for IBRE is debatable even among key researchers. Kruschke, an advocate of Attentional Theory, showed that in his 2001 study that IBRE is not explained by the eliminative inference approach. Winman et. al. (2003), advocates of the eliminative inference approach, conducted another study to refute Kruschke's Attentional theory explanation of IBRE by conducting simulation using Kruschke's 1996 experimental data on one of Kruschke's categorization network models, the ADIT model. For this model, Winman et al. (2003) obtained the proportion of correct classification by varying the ADIT network model's weight learning rates and the attentional shifting rates. Using the ADIT simulation categorization model, Winman et al. (2003), tested the assumption that rapid attention shifts accelerate learning and that IBRE is a direct consequence of an asymmetry between associations or relationships $PR \rightarrow R$ and $PC \rightarrow C$. The conclusion from the implementation of the ADIT network model was that rapid attention ship does not accelerate learning (i.e., proportions of correct classifications) on the IBRE design, that the most rapid learning was obtained when no attentional shift occurred, and that relationships predicted by Attentional Theory in the ADIT model do not support IBRE but there was support attributed to the direction of base rates. (Winman, et al., 2003). This ADIT simulation also supported the idea that higher frequency of occurrence of a category predicts a preference for the common category C rather than the rare category R when confronted with two conflicting stimuli PR and PC.

Kruschke, (2003) responded to Winman et al. 2003's study by conducting a study on his own on the same data that Winman et al. (2003) used. Essentially, Kruschke (2003) claimed that the ADIT network model was not adequate to assess IBRE because it is a model that *does not learn* its shifts of attention or its attention shift parameters do not change after every trial. Kruschke (2003) repeated Winman et al.'s 2003 study again on Kruschke's 1996 experimental data with another simulation categorization model, the EXIT network model. According to Kruschke (2003), the EXIT network model incorporates not only attentional shifting rates but also attentional learning rates. When these parameters are adjusted in this network model, the results indicate that attentional shifts do produce inverse base rate effects (IBRE) in the model, regardless of whether the shifts are learned (Kruschke, 2003). Thus, based on his results, Kruschke (2003) concluded that Attentional Theory is a possible explanation for IBRE. In summary, conflicting results exist as to what exactly causes the IBRE phenomenon.

4.7 Recent Trends

According to Ashby and Maddox (2005) the changes that have evolved in the recent years in the field of human categorization has centered, both theoretical and empirical, on two main issues: a) the field embracing the cognitive neuroscience field and b) the human category learning processes being mediated by multiple distinct learning systems instead of a single learning system. In the neuroscience field, patients with frontal lobe lesions, patients with a disease of the basal ganglia (e.g. Parkinson's or Huntington's disease), and patients with amnesia whose amnesia was caused by damage to the medial temporal lobes, are used in categorization learning studies. One tool that is used in these categorization studies is the Wisconsin Card Sorting Test (WCST) where the stimuli are cards containing different geometric patterns. According to Ashby and Maddox (2005), in neuropsychological assessments, practically all

category learning tasks are rule based. In rule based tasks, the categories are learned via some explicit reasoning process and dimensions can be combined using conjunctions such as “and” to formalize a rule. Decisions are made about each dimension before they are combined. (This is different from information-integration tasks in that all dimensions are combined *before* any decision is made.) Neuropsychological studies indicate that damage to the brain’s frontal-striatal circuits (frontal lobe lesions) and to the brain’s basal ganglia affect rule-based category learning but damage to the brains’ medial temporal lobe do not in rule based category learning. (Janowsky et al, 1989, Leng and Parking, 1988). Other neuroimaging data on brain damaged patients also support the idea that when patients are performing rule based tasks, certain parts of the brain (the prefrontal cortex, the head of the caudate nucleus, the anterior cingulate) become activated and damage to these structures caused impairment on rule based tasks (Ashby and Maddox ,2005).

A model that has a neuropsychological basis and is used to explain that human categorization learning occurs via different systems is the Competition between Verbal and Implicit Systems (COVIS) model. Briefly, the COVIS model is used to explain that there are two separate category learning systems (explicit and implicit) that humans use that are linked through two different brain paths.

The COVIS model assumes that rule based learning is mediated by an explicit category learning system that relies on working memory and executive attention. For explicit learning, rules are stored in a person’s working memory until they need to be tested. When feedback prompts for a different rule, a new rule is activated through two different processes: a) the selection of the new rule and b) the attention switch to the new rule. According to Ashby et al. (1998, 1999), the selection operation is mediated cortically, by the anterior cingulate and

possibly also by the prefrontal cortex of the brain, and that switching is mediated by the head of the caudate nucleus of the brain. The probability of selecting the new rule depends on its reinforcement history, the person's tendency to select new hypothesis, and the person's tendency to perseverate (Ashby et. al., 1999).

Another recent idea on human categorization learning that is supported by Ashby and Maddox (2005) research, one of the most cited works in human categorization learning, is the idea that how humans learn categories depend on the type of task that they are assigned to learn. Ashby and Maddox (2005) research describes four types of category learning tasks: rule-based tasks, information-integration tasks, prototype distortion tasks, and weather prediction to test the main idea. For rule based tasks, the categories are supposedly learned via some explicit reasoning process. Example of a rule based task is, "If the animal barks and has four legs, then respond 'dog'." Often there is an optimal rule that will maximize the categorization accuracy but it does not mean it is always used. Information integration category learning tasks required the integration of information from two or more stimulus components (e.g., determining whether an X-ray shows a cancerous tumor is an information integration task that requires the integration of many cues and strategies.). Prototype distortion tasks are tasks that involve different exemplars, versions of the category prototype. If the prototype is five dots, then the arrangements of these five dots in different configurations are considered exemplars. Weather prediction tasks require the probability of outcome occurrence to be determined for categorization purposes.

For each of these tasks, categorization accuracy can depend on many factors. Changes in categorization procedures can affect the categorization accuracy of information-integration tasks and but not of rule based tasks. Ashby et al. (2003) research observed that response time to categorization was affected for information-integration tasks but not for rule-based tasks by

changing the categorization procedures. Short delays in feedback categorization information were also shown to affect rule based categorization tasks but not information-integration tasks (Maddox et al, 2004). This indicated that feedback processing requires attention and effort for rule based categorization learning but not for information-integration learning. The linear and non-linear combination of dimensions in information integration tasks can affect the accuracy of categorization. According to Ashby and Maddox (2005), it is much more difficult to learn nonlinear separable categories than linear separable categories. Information integration studies that have been conducted on brain impaired patients have suggested that Parkinson's disease subjects are impaired in information-integration tasks, but only if the category structures are nonlinearly separable and complex (Ashby and Maddox , 2005). Lastly, for prototype distortion tasks, for some unknown reason patients who have shown to have problems with rule based tasks and information integration tasks do not show problems with prototype distortion tasks. This includes patients with Parkinson's disease, Schizophrenia or Alzheimer's disease (Ashby and Maddox, 2005).

4.8 Conclusion

In conclusion, the field of human categorization is a vast field that has expanded over the years in many directions. In this dissertation, an attempt was made to introduce the reader to the issues that currently exist in the human categorization field. The quest for how humans actually perform categorization is an ongoing endeavor. In this particular study, human categorization was performed in a quality context. The focus of this study was not to investigate how humans actually perform categorization but to assess the performance of human classification relative to two computer algorithms: latent semantic analysis and latent dirichlet allocation.

CHAPTER 5

THEORY DEVELOPMENT

5.1 Introduction

The ability for humans to categorize is considered to be a foundation of cognition in human categorization literature (Arterberry and Bornstein, 2002). Many of the rational models of cognition that have been developed for categorization (Anderson, 1990; Ashby & Alfonso-Reese, 1995) reveal connections between human behavior and information systems that solve similar problems (Griffiths, et al., 2007). Human categorization is an area of the cognitive science field that can be shown to be linked to information system field. In this dissertation, it will be shown that the foundations of the computer algorithm LDA and its predecessor LSA are highly linked to theories in human categorization that are found in the cognitive science field.

5.2 Latent Semantic Analysis (LSA) as a Proposed Theory of Knowledge

As mentioned in Chapter 2, LSA is a text mining technique that replaces the original term-document matrix by a filtered term-document matrix which serves as a better representation of the semantic content in terms and documents. In LSA, a document representation is a vector of words in vector space that can be compared to another vector for the retrieval of similar documents. By simplifying the document representation in the LSA component space, LSA facilitates the retrieval of similar documents even when they do not literally share the same terms, but they contain synonym terms or they cover a similar topic in an abstract manner. LSA, which is highly used in topic extraction, has been linked to the acquisition of human knowledge. Laundaer and Dumais (1977) mentioned that children acquire human knowledge as they improve their reading skills and learn to associate more similar words with time. Their learning

knowledge is usually measured through synonym tests. Landauer and Dumais (1997) research on LSA evaluated the performance of LSA on similar synonym tests and discovered that the performance of LSA on these tests depend on the dimensionality used in the LSA model. Based on the results achieved, Landauer and Dumais (1977) concluded that LSA can be used as a theory to partly explain knowledge acquisition in humans. Based on this existing literature, comparing LSA to human categorization performance as performed in this study is relevant.

5.3 Latent Dirichlet Allocation (LDA) as a Proposed Model of Human Cognition

The latent dirichlet allocation (LDA) method evolved from the development of latent semantic analysis (LSA). As an intermediate step, pLSA, was introduced in 1999 by Thomas Hofman, designed to better deal with polysemy issues that LSA could not effectively address (Hofman, 2001). pLSA is an extension of LSA and it differs from LSA in that it has a statistical foundation. The basis for pLSA is the statistical model called the aspect model. The assumptions for the aspect model are: a) the “bag of words” assumption where the order of appearance of words in a document do not matter and b) the probability of a word being in a document $p(w,d)$ is generated independently. A detailed description of LDA is found on section 2.4.4.

As with LSA, LDA has been mostly usually used for topic extraction. The performance of LDA in synonyms tests can also be evaluated in the same manner as Landauer and Dumais (1977) did with LSA in an attempt to measure how relevant is LSA to human performance in knowledge acquisition. In evaluating these two methods to explain human learning performance, the question that often arises is how humans learn better. In particular for categorization, do humans learn better via the principal components method (LSA) or via the statistical model (LDA)? Which model can serve better to explain human cognition? In this dissertation, one

important assumption of LDA is the aspect model assumption that is used to derive key probabilities in LDA. In this chapter, the aspect model assumption of LDA will be shown to be linked to the human categorization theories and thus, the potential for LDA to serve as a model of human cognition exists. In particular, two main human categorization theories that will be shown to be linked to the development of LDA are the Prototype theory and Exemplar theory.

5.4 Prototype and Exemplar Theories

Prototype theory explains that categorization is made in reference to a central category member, a prototype. A category can have different exemplars, versions, of the central prototype. When an unfamiliar stimulus is encountered, it is assigned to the category with the most similar prototype (Homa et al., 1981; Posner & Keele 1968, 1970; Smith & Minda, 1998). According to Ashby and Maddox (2005), “Prototype theory assumes a category is represented as a prototype, and stimuli are categorized by comparing them to the prototypes of each contrasting category.”

Exemplar theory also assumes that there are many versions (exemplars) that exist within each category. When an unfamiliar stimulus is encountered, it is compared to *every* exemplar that exists within a category. The stimulus is assigned to the category where its similarity is the highest in terms of the number of exemplars within a category (Brooks, 1978; Estes, 1986, 1994; Hintzman, 1986; Lamberts, 2000; Medin & Schaffer, 1978; Nosofsky, 1986). To better illustrate the difference between these two theories, Ashby and Maddox (2005) provided an analogy of how prototype theory and exemplar theory differ by using the solar system, in which the sun is the central prototype and the planets are distortions (exemplars) of the central prototype. The probability of an individual responding “A” in an “(A, not A)” category task can change with the position of the stimulus in the solar system space model. Under prototype theory, the probability

of responding “A” depends on the distance between the stimulus and the central prototype, the sun. With exemplar theory, the probability of responding “A” depends on the sum of the distance between the stimulus and all of the exemplars, the planets. Graphically, the two theories are illustrated in Figures 5.1 and 5.2.

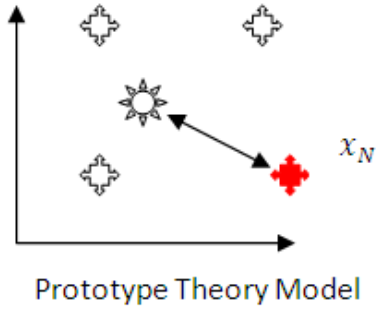


Figure 5.1. Prototype Theory Model.

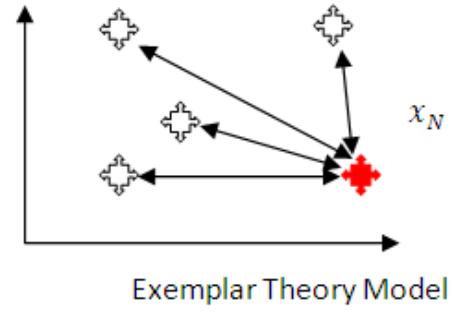


Figure 5.2. Exemplar Theory Model.

Griffiths et al. (2008) formalizes the difference between human categorization exemplar models and the prototype models in the following way: Given a set of $N-1$ stimuli with features $\mathbf{x}_{N-1} = (x_1, x_2, x_3, \dots, x_{N-1})$ and category labels $\mathbf{y}_{N-1} = (y_1, y_2, y_3, \dots, y_{N-1})$, the probability that a new stimulus N with features x_N is assigned to category j is given by:

$$P(y_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \frac{\eta_{N,j} \beta_j}{\sum_y \eta_{N,j} \beta_j} \quad (34)$$

Where $\eta_{N,j}$ is the similarity of the stimulus x_N to category j and β_j is the response bias for category j . The key difference between the two models is how the $\eta_{N,j}$, the similarity of the stimulus x_N to category j , is computed. For the exemplar theory models, $\eta_{N,j} = \sum_{i|y_i=j} s_{N,i}$ where $s_{N,i}$ is a measure of the similarity between stimuli x_N and x_i . For the prototype theory

models, $\eta_{N,j} = s_{N,P_j}$ where s_{N,P_j} is a measure of the similarity between stimulus N and the prototype P_j where

$$P_j = \frac{1}{N_j} \sum_{i|y_i=j} x_i \quad (35)$$

where N_j = number of instances of that category.

5.5 Linking the Prototype and Exemplar Theory

Vanpaemel et al. (2005) were able to show that the Prototype Theory and Exemplar Theory were linked through the introduction of clustering of instances. Vanpaemel et al. (2005) formalized a set of models by portioning instances of each category into clusters where the number of cluster for category K_j can range from 1 to N_j where N_j is the total number of instances of category j and each cluster k is represented by a prototype $p_{j,k}$ for category j . The similarity of stimulus N to category j , $\eta_{N,j}$, is defined to be:

$$\eta_{N,j} = \sum_{k=1}^{K_j} s_{N,P_{j,k}} \quad (36)$$

If $K_j=1$, this is equivalent to the prototype model. If $K_j=N_j$, this is equivalent to the exemplar model. Or in simpler words, if all the instances labeled with category j go under one cluster, then this is the prototype model. If all the instances labeled with category j go each under separate clusters, then this is the exemplar model. Thus, the size of the cluster determines whether the model supports Prototype theory or Exemplar theory. Clustering of instances in real life application has some rationality. For example, groups of males and females with different features can be further subdivided into clusters, each signifying a different race. For each cluster, a single prototype can be defined.

Ashby and Alfonso-Reese (1995) connected the similarity $\eta_{N,j}$ in the prototype and exemplar models to Bayesian probabilities. They identified the similarity $\eta_{N,j}$ as the probability of generating an item $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$ from a category j and the bias term β_j as the prior probability of category j , $P(y_N=j|\mathbf{y}_{N-1})$. In this manner, equation 34 was transformed using Bayes' Rule into the following equation whose probability terms could be obtained through sampling from a mixture.

$$P(y_N = j | \mathbf{x}_N, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \frac{P(x_N = j | y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})^* P(y_N = j | \mathbf{y}_{N-1})}{\sum_y P(x_N = y | y_N = y, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})^* P(y_N = y | \mathbf{y}_{N-1})} \quad (37)$$

5.6 Generic Mixture of Classification

The clustering of instances or stimuli leads to the generic mixture of classification as explained by Rosseel (2002). Rosseel's (2002) extended Vanpaemel et al.'s (2005) work to include clusters across many categories as opposed to just within categories. According to Rosseel (2002), many models of categorization estimate the likelihood that a stimulus x belongs to one of several categories C_k , where $k=1, \dots, k$ (Ashby & Alfonso-Reese, 1995). This is defined by $P(x/C_k)$. According to Rosseel (2002), the Generic Mixture Model of Categorization makes three assumptions:

- a) A specific exemplar x_n is represented perceptually by a random vector having a multivariate distribution $p(x/n)$ centered on x_n and with a covariance matrix Σ_{pn} .
- b) The unconditional probability density function of a set of exemplars (belonging to K categories) is modeled as a finite mixture distribution.

$$p(x) = \sum_{j=1}^J P(j)p(x|j) \quad (38)$$

Here, J is the number of mixture components used in the mixture model; $P(j)$ denotes the unconditional mixture proportion and $p(x|j)$ are the individual component densities.

- c) The probability density function of category C_k is modeled as a finite mixture distribution sharing the same mixture components $p(x|j)$ of the unconditional mixture distribution $p(x)$.

$$p(x|C_k) = \sum_{j=1}^J P(j|C_k)p(x|j) \quad (39)$$

In categorization, *prior* to assigning exemplar x to a category C_k , the probability $p(x|C_k)$ needs to be computed for all $k=1 \dots N_k$ categories where $N_k =$ total number of categories. The exemplar is assigned to the category k for which $p(x|C_k)$ is the highest. As an illustration, the graphical picture shown in Figure 5.3 is used.

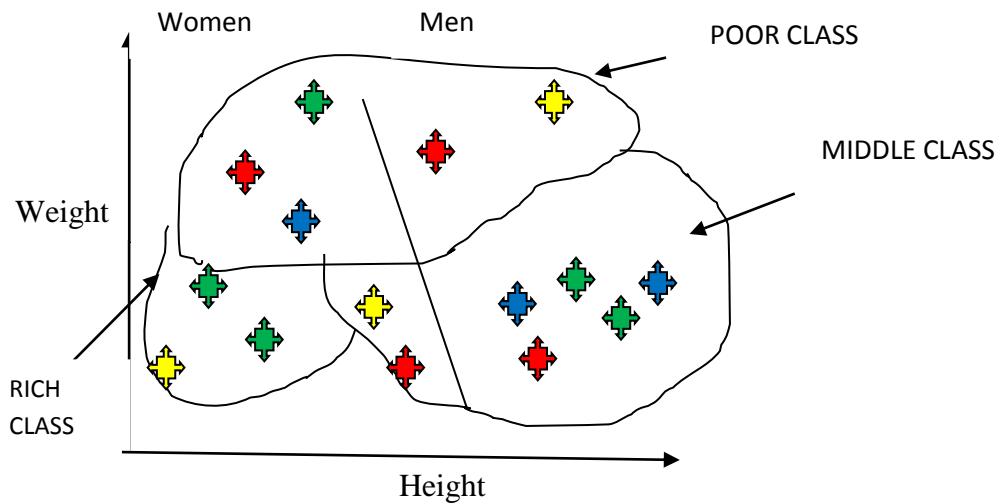



Figure 5.3. General mixture classification example.

Let a person with features x be , the question is: should this person be assigned to category Women or to Category Men?

Computing $p(x|C_k)$ for $C_k = \text{Women}$, the following relation yields:

$$p(x|C_k) = \sum_{j=1}^J P(j|C_k)p(x|j) \quad (40)$$

$$p(x|C_k) = P(\text{poor/women})P(x/\text{poor}) + P(\text{middle class/women})p(x/\text{middle class}) + \\ p(\text{rich/women})p(x/\text{rich}) = (3/5)*(2/5) + (2/7)(2/7) + 1(0) = 6/25 + 4/49 = .322$$

For $C_k = \text{Men}$

$$p(x|C_k) = P(\text{poor/men})p(x/\text{poor}) + p(\text{middle class/men})p(x/\text{middle class}) + p(\text{rich/men})p(x/\text{rich}) \\ = (2/5)(2/5) + (5/7)(2/7) + (0)(0) = 0.360$$

According to the status of the mixture model, this person would have to be assigned to the category Men.

The computation of the $p(x|C_k)$ is similar to the computation of a word w_i for document d in the aspect model used in LDA.

$$P(w_i|d) = \sum_{j=1}^T P(w_i|z_i=j)P(z_i=j|d) \quad (41)$$

The analogy between probabilities computed by generic mixture of categorization and the aspect model are the following: instance x is analogous to word w_i , cluster j is analogous to topic z , and document d is analogous to category C_k . In this form, the theories of human categorization learning are linked to the development of the LDA text mining method. Using LDA as a classifier and comparing its performance to human classification is relevant. In LDA, we sample from a mixture of words to determine the probability that a word belongs to a document. In the generic mixture of categorization, humans compute the probability that a

particular instance would belong to a category based on the existing mixture of instances, *prior* to assigning that instance to a category. Human subjects determine the probability of an instance belonging to a category by sampling from the existing mixture of instances. Thus, in the generic mixture of categorization and in the aspect model, sampling is done from a mixture. In the aspect model, the sampling is performed to retrieve a word from the mixture and compute the probability for that word belonging to a document. In the generic mixture of categorization, the goal is to sample from the existing mixture and determine the probability that a similar item (to the incoming item that requires classification) exists in the mixture for each category. Then based on the computed probabilities of this item existing in each category, the new item is assigned to a specific category. The big difference in the two methods is that in the generic mixture of categorization, the incoming item was never part of the mixture when sampling was going on, but after sampling was finished, it became part of the mixture. In this manner, the development of LDA is linked to the theories of human categorization. Thus, comparing the performance of LDA algorithms to human categorization performance has a common statistical basis.

Based on the LDA and human categorization having similar theoretical statistical foundations, we state the following proposition:

P1: Human category performance is closer to LDA performance than LSA performance.

In the next chapter, we formulate testable hypotheses related to this proposition.

CHAPTER 6

RESEARCH METHODOLOGY

6.1 Overview

In order to assess classification performance by humans, LDA, and LSA, a data collection study was designed. The study aimed at generating a set of documents with known classification labels by a group of subjects. The study consisted of three stages: the generation of documents (i.e., customer comments), the human categorization of these documents, and the computer categorization of these documents using the latent dirichlet allocation (LDA) method and the latent semantic analysis (LSA) method.

6.2 First Stage: Comment Generation

In the first stage, customer comments were generated via an online survey that was administered to a group of business students enrolled in six beginning business statistics classes from a southwestern public university who acted as customers of a pizza restaurant. The comments were generated in response to a service quality issue that is common in a pizza restaurant setting-how to achieve customer satisfaction. As a starting guideline for the generation of these comments, a fish-bone diagram was created through a brainstorming session to determine the major and minor causes that would lead to customer dissatisfaction. The fish-bone diagram is a tool that is used often in quality assurance to analyze cause and effect situations. In this case, quality was defined as “meeting or exceeding customer’s expectations” (Evans and Lindsay, 1999) with a pizza restaurant. The objective of the fish-bone diagram was to illustrate the causes that will cause the main effect of being “customers’ dissatisfaction” to occur. Based on the developed fish bone diagram illustrated below, 21 specific causes

(subtopics) were identified as the cause for customers' dissatisfaction. These subtopics were rearranged into four main cause categories: Food Condition, Food Processing Methods, Facility, and Customer Service. Figure 1 illustrates the fish-bone that was created as a result of the brainstorming session.

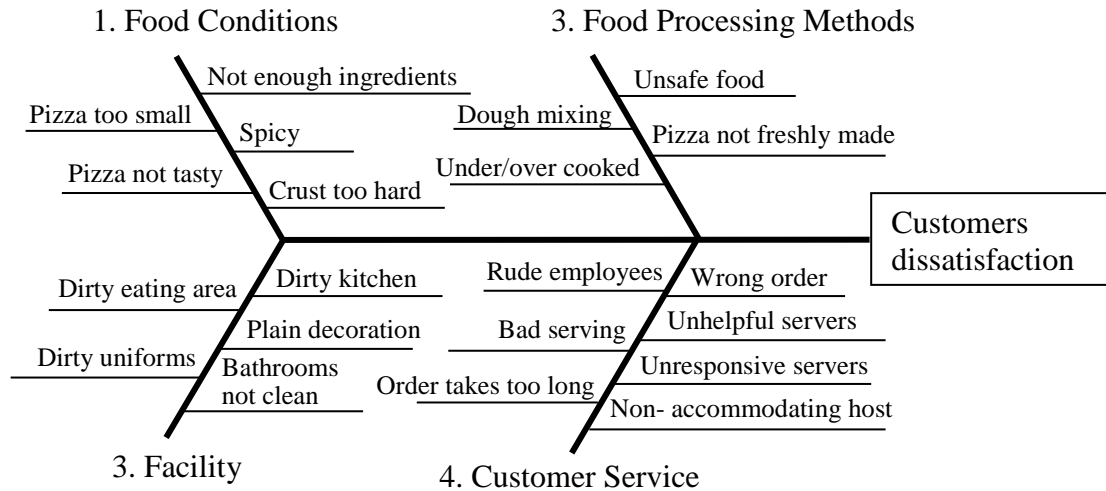


Figure 6.1. Fish-bone diagram.

- Major cause category 1: Food Condition
- Major cause category 2: Food Processing Methods
- Major cause category 3: Facility
- Major cause category 4: Customer Service

Using this fish-bone diagram as a guideline, an online survey was developed where participants had an opportunity to make comments on 63 randomly selected questions or make three random comments on each of the 21 subtopics from the fish-bone diagram. The participants accepted the terms of the university's Institutional Review Board prior to beginning the survey. Appendix B lists the survey format that was used for this stage of the research. For each question, a participant was asked to provide a description statement. This description was

to include (1) an everyday example of the problem cause as it would be worded by a complaining customer, and (2) the specific cause and (3) the major cause category.

To generate a large variety of comments, six surveys were administered to six different business classes during the Fall 2010 semester. A total of 6000+ comments were generated from 300+ students. The online survey also recorded the time that a participant spent completing the survey. This time was then used to determine which of these comments were deemed invalid and thus needed to be discarded. Essentially, if a participant did not spend a reasonable amount of time on the survey, then his/her comments were deemed to be invalid. Out of the remaining comments, 1008 comments were used for the human categorization stage of the experiment.

To study how human participants classify comments, a classification experiment was designed. Classification is a form of supervised learning. The process of assigning pre-defined category labels to a new database tuple based on the analysis of a labeled training data set is called classification (Han and Kamber, 2006). In classification, the data set has labeled data points where each label, ω_k , belongs to a set of c classes or $\omega_k \in \{1, 2, \dots, c\}$ where c is the number of classes. Usually, part of this data set, the training set, is usually used to develop a classifier, Φ , that will be used to later predict categories on new data points. Although there is not a consistent definition for a classifier, a classifier, Φ , can be an algorithms, a decision rule, a mathematical formula and/or any other techniques that can be used to discriminate between classes of patterns. Classifiers can be multi-class or dual class. The type of classifier used depends on the character of the data, the number of classes, the learning algorithm that created the classifier and the validation procedures. In this study, the classifiers were humans in the second stage and they were computer algorithms (LSA and LDA) in the third stage.

6.3 Second Stage: Human Categorization

The classification study consisted of a second online survey that was developed to incorporate the customer comments from the first survey. Prior to filling out the survey, students had to accept the terms of the university's Institutional Review Board. Appendix C lists the format of the survey that was used for this stage. In this study, sixteen online surveys were created. Each consisted of 63 comments, representing randomly three comments from each of the 21 subtopics. These surveys were administered to business students during the last days of the Fall 2010 semester.

For these sixteen surveys, the inter-rater reliability was computed. According to Kilem Li Gwet (2008), "the inter-reliability is regularly used in medical and social research to evaluate the reliability of rating systems." In the medical field, psychiatrists may develop a protocol to classify mental diseases on a patient that is used by nurses who act as raters. Their level of agreement is desired for the correct diagnosis of the patient's disease. Two common measures to determine the inter-rater reliability of raters are Gwet's AC1 Statistic as well as Fleiss' generalized Kappa (Gwet,2008). For these surveys, the AC1 and the Kappa statistics were computed. Gwet (2008) showed in his research that his AC1 statistic is highly related to the Kappa statistic. The results indicated that based on Landis and et al. (1977), the inter-rater strength of agreement was substantial based on the Kappa statistic for fourteen out of the sixteen surveys. Based on these results, survey number 2 and number 10 were discarded as their inter-rater reliability measures showed less than a substantial strength of agreement. The remaining surveys were used for the analysis of this research and the inter-rater reliability values are given below for these surveys.

Table 6.1:

Inter-Rater Reliabilities

Survey	Raters	AC1	Kappa
v01	3.000	0.700	0.699
v02	2.000	0.584	0.584
v03	2.000	0.800	0.800
v04	2.000	0.733	0.733
v05	4.000	0.831	0.830
v06	3.000	0.678	0.677
v07	2.000	0.867	0.866
v08	2.000	0.767	0.765
v09	3.000	0.850	0.850
v10	3.000	0.556	0.555
v11	2.000	0.701	0.700
v12	3.000	0.867	0.866
v13	3.000	0.861	0.861
v14	3.000	0.784	0.783
v15	3.000	0.889	0.889
v16	4.000	0.853	0.853

6.4 Management Interpretation of Comments

Although the focus of this study is to compare the performance of human classification to the classification of two text mining techniques, latent dirichlet allocation (LDA) and latent semantic analysis (LSA), another research question arose in the evaluation of the gathered human categorization of these sixteen online surveys. The question that arose is that when customers articulate their quality concerns as being the source of their dissatisfaction with the service, how accurate are these concerns being interpreted correctly by management? As established by Iglesias (2004), the voice of the customer not only judges the quality attributes of a service operation but also comes with preconceptions and expectations. Thus, the correct interpretation of complaints can affect service operations and it was recognized that the analysis

of this data also had implications for the management of a service operation and these implications are detailed in Appendix D through a conference paper that will be presented on November 21, 2011 at the 2011 Annual Decision Science institute Annual Conference in Boston, Massachusetts. Briefly, this paper analyzes these sixteen surveys to determine the accuracy of management interpretations of customer complaints (voice of the customer) of a service operation.

6.5 Classification Performance Measures

To analyze the collected data for classification, the performance classification measures macro- F_1 scores and micro- F_1 scores were used. According to Tang et al. (2009), given a test data $X \in \mathbb{R}^{N \times M}$, let $y_i, \hat{y}_i \in \{0,1\}^K$ be the true label set and the predicted label set for instance x_i . The Macro- F_1 score is defined as the F_1 score averaged over all K categories.

$$Macro - F_1 = \frac{1}{K} \sum_{k=1}^K F_1^k \quad (42)$$

To determine each individual F_1^k score for each category K , the precision P^k and recall R^k need to be computed. P^k is the proportion of correct predicted responses from all the predicted scores for category k . R^k is the proportion of correct responses from all actual responses for category k .

$$P^k = \frac{\sum_{i=1}^N y_i^k \hat{y}_i^k}{\sum_{i=1}^N \hat{y}_i^k} \quad \text{and} \quad R^k = \frac{\sum_{i=1}^N y_i^k \hat{y}_i^k}{\sum_{i=1}^N y_i^k} \quad (43)$$

The F_1^k measure is defined as

$$F_1^k = \frac{2 P^k R^k}{P^k + R^k} \quad (44)$$

The micro- F_1 score is defined as:

$$Micro - F_1 = \frac{2 \sum_{k=1}^K \sum_{i=1}^N y_i^k \hat{y}_i^k}{\sum_{k=1}^K \sum_{i=1}^N y_i^k + \sum_{k=1}^K \sum_{i=1}^N \hat{y}_i^k} \quad (45)$$

In addition to the macro-F₁ and micro-F₁ scores, the classification experiments were evaluated for accuracy of classification.

6.6 Analysis of First Set of Surveys

Analysis of the first set of surveys was done to develop the conference paper that was submitted to the Decision Science Institute conference. The results include the accuracy, Macro-F₁ and Micro-F₁ classification scores for the subtopics.

Table 6.2

First Surveys Subtopics Classifications Scores (44 participants)

ResponseSet	Student	Average Macro F ₁	Micro F ₁ Score Per	Accuracy
v01	1	0.5154	0.5397	0.5397
v01	2	0.6690	0.6825	0.6825
v01	3	0.6646	0.6667	0.6667
v02	4	0.4387	0.4800	0.4762
v02	5	0.6726	0.6667	0.6667
v03	6	0.5832	0.5873	0.5873
v03	7	0.6257	0.6349	0.6349
v04	8	0.5836	0.6190	0.6190
v04	9	0.5966	0.5873	0.5873
v05	10	0.6721	0.6667	0.6667
v05	11	0.6619	0.6667	0.6667
v05	12	0.6816	0.6825	0.6825
v05	13	0.5302	0.5556	0.5556
v06	14	0.5295	0.5397	0.5397
v06	15	0.5520	0.5556	0.5556
v06	16	0.6249	0.6190	0.6190
v07	17	0.6384	0.6349	0.6349
v07	18	0.6935	0.6984	0.6984
v08	19	0.6130	0.6349	0.6349
v08	20	0.6288	0.6508	0.6508

(table continues)

Table 6.2 (continued)

v09	21	0.6121	0.6349	0.6349
v09	22	0.6422	0.6349	0.6349
v09	23	0.6779	0.6720	0.6667
v10	24	0.5683	0.5714	0.5714
v10	25	0.6499	0.6508	0.6508
v10	26	0.4382	0.4444	0.4444
v11	27	0.5245	0.5440	0.5397
v11	28	0.6821	0.6825	0.6825
v12	29	0.6423	0.6349	0.6349
v12	30	0.6389	0.6508	0.6508
v12	31	0.5746	0.5873	0.5873
v13	32	0.6039	0.6190	0.6190
v13	33	0.6384	0.6349	0.6349
v13	34	0.5991	0.6190	0.6190
v14	35	0.5915	0.5873	0.5873
v14	36	0.6141	0.6190	0.6190
v14	37	0.6138	0.5968	0.5873
v15	38	0.6582	0.6667	0.6667
v15	39	0.6721	0.6667	0.6667
v15	40	0.6313	0.6179	0.6032
v16	41	0.6154	0.6508	0.6508
v16	42	0.6197	0.6508	0.6508
v16	43	0.6952	0.6825	0.6825
v16	44	0.6626	0.6508	0.6508

The macro- F_1 scores and micro- F_1 Scores for the main categories are:

Table 6.3

First Surveys Main Categories Classification Scores (44 participants)

ResponseSet	Average Per Student Macro F_1 Score	Micro F_1 Score Per student	Accuracy
v01	0.8107	0.4127	0.8254
v01	0.9472	0.4762	0.9524
v01	0.9444	0.4762	0.9524
v02	0.2850	0.1600	0.3175
v02	0.2998	0.1587	0.3175
v03	0.9258	0.4683	0.9365
v03	0.9271	0.4683	0.9365
v04	0.8846	0.4524	0.9048
v04	0.9007	0.4524	0.9048
v05	0.9666	0.4841	0.9683

(table continues)

Table 6.3 (continued)

v05	0.9811	0.4921	0.9841
v05	0.9811	0.4921	0.9841
v05	0.9625	0.4841	0.9683
v06	0.8798	0.4444	0.8889
v06	0.8390	0.4286	0.8571
v06	0.9114	0.4603	0.9206
v07	0.9330	0.4683	0.9365
v07	0.9858	0.4921	0.9841
v08	0.9629	0.4841	0.9683
v08	0.9285	0.4683	0.9365
v09	0.9285	0.4683	0.9365
v09	0.9629	0.4841	0.9683
v09	0.9411	0.4720	0.9365
v10	0.9814	0.4921	0.9841
v10	1.0000	0.5000	1.0000
v10	0.7233	0.3651	0.7302
v11	0.9204	0.4640	0.9206
v11	0.9625	0.4841	0.9683
v12	0.9629	0.4841	0.9683
v12	0.9629	0.4841	0.9683
v12	0.9344	0.4683	0.9365
v13	0.9279	0.4683	0.9365
v13	0.9311	0.4683	0.9365
v13	0.8975	0.4524	0.9048
v14	0.9094	0.4603	0.9206
v14	0.9330	0.4683	0.9365
v14	0.9488	0.4758	0.9365
v15	0.9814	0.4921	0.9841
v15	0.9814	0.4921	0.9841
v15	0.9383	0.4715	0.9206
v16	0.9444	0.4762	0.9524
v16	0.9333	0.4683	0.9365
v16	0.9814	0.4921	0.9841
v16	0.9629	0.4841	0.9683

To compare the degree to which participants categorized the four broad quality categories, a one-way ANOVA was performed on the macro F-scores for the four main categories: M1= Food Condition, M2=Food Processing, M3=Facility and M4= Customer Service. The results indicated that highly significant differences existed among the four categories (p-value < 0.001).

Table 6.4

ANOVA of Main Categories

One-way ANOVA: M1, M2, M3, M4

Source	DF	SS	MS	F	P
Factor	3	0.23127	0.07709	31.1	0
Error	152	0.37682	0.00248		
Total	155	0.60809			

S = 0.04979 R-Sq = 38.03% R-Sq(adj) = 36.81%

Table 6.5

95% Confidence Intervals for Main Categories

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev	-----+-----+-----+-----+
M1	39	0.907	0.058	(---*---)
M2	39	0.890	0.069	(---*---)
M3	39	0.973	0.030	(---*---)
M4	39	0.976	0.030	(---*---)

-----+-----+-----+-----+

0.900 0.930 0.960 0.990

Pooled StDev = 0.04979

The results indicate highly significant differences among the four categories (p-value < 0.001).

To further investigate those differences we also performed a Tukey test for multiple comparisons, with a family error rate of 5% and a corresponding individual category error rate of 1.03%.

Table 6.6

Tukey Method Grouping of Main Categories

Grouping Information Using Tukey Method

	N	Mean	Grouping
M4	39	0.976	A
M3	39	0.973	A
M1	39	0.907	B
M2	39	0.890	B

Means that do not share a letter are significant different.

The results indicate that the four categories could be grouped into two distinct groups: {M1, M2} and {M3, M4}.

Similarly, an ANOVA was performed to on the subtopics. The analysis below compares the 21 specific categories listed below:

- S01= Food Condition: Pizza is not very tasty.
- S02= Food Condition: Pizza size is too small.
- S03=Food Condition: Quantity of pizza ingredients such as cheese or toppings is too small (restaurant) is being too stingy).
- S04= Food Condition: Pizza has either too much or too little amount of spices.
- S05= Food Condition: Pizza crust is too hard.
- S06= Food Processing Methods: Pizza is under or overcooked.
- S07= Food Processing Methods: Pizza was not safe to eat (possible contamination).
- S08= Food Processing Methods: Pizza does not look freshly made (cold pizza)
- S09= Food Processing Methods: Pizza dough was not mixed well (has lumps inside).
- S10=Facility: The restaurant’s entire waiting area looks dirty.
- S11=Facility: Employee uniforms are dirty.
- S12=Facility: Kitchen area looks dirty, disorganized, messy.
- S13=Facility: Environment is just too plain, too undecorated.
- S14=Facility: Bathrooms are not clean.
- S15=Customer Service: Employees are impolite or rude.
- S16=Customer Service: When I need help, it takes too long for the server to notice.
- S17=Customer Service: Server is not helpful while I try to make my order decisions.
- S18=Customer Service: Host is not willing to accommodate my seating preference.
- S19=Customer Service: My order takes too long to be served.
- S20=Customer Service: Server brings the wrong order (wrong crust, wrong toppings,
- S21=Customer Service: Pizza is not served appropriately (plates, napkins, knives, etc.).

Table 6.7

ANOVA of Subtopics (p-value <0.001)

One-way ANOVA: S01, S02, S03, S04, S05, S06, S07, S08, S09, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21

Source	DF	SS	MS	F	P
Factor	20	6.944	0.347	10.840	0
Error	798	25.564	0.032		
Total	818	32.508			

S = 0.1790 R-Sq = 21.36% R-Sq(adj) = 19.39%

The results indicated that highly significant differences existed among the 21 subtopics.

Table 6.8

95% Confidence Intervals for Subtopics

Individual 95% CIs For Mean Based on Pooled StDev					
Level	N	Mean	StDev	-----+-----+-----+-----+-----	
S01	39	0.5739	0.2935	(--*--)	
S02	39	0.9017	0.1526		(---*---)
S03	39	0.8774	0.1267		(--*--)
S04	39	0.8310	0.2372		(--*--)
S05	39	0.7570	0.2913	(---*---)	
S06	39	0.7527	0.1337	(---*---)	
S07	39	0.8063	0.2078		(---*---)
S08	39	0.8568	0.1967		(---*---)
S09	39	0.9339	0.1182		(--*--)
S10	39	0.9513	0.0997		(--*--)
S11	39	0.9337	0.1835		(---*---)
S12	39	0.9131	0.1334		(---*---)
S13	39	0.9685	0.0693		(---*---)
S14	39	0.9392	0.1065		(---*---)
S15	39	0.8498	0.1354	(---*---)	
S16	39	0.7471	0.2778	(---*---)	
S17	39	0.8625	0.1785		(---*---)
S18	39	0.9465	0.1245		(---*---)
S19	39	0.8814	0.1427		(---*---)
S20	39	0.9084	0.1916		(---*---)
S21	39	0.9328	0.1223		(---*---)

-----+-----+-----+-----+-----

0.60 0.75 0.90 1.05

Pooled StDev = 0.1790

The results indicate that there are substantial differences on the classification of these specific subtopics by the participants. The results also indicate that subtopic number S01 stands out as an outlier from the rest of the subtopics.

6.7 Human Classification Experiment

It was recognized that additional data was needed for the classification experiment. The classification-by-humans stage was repeated again with the creation of twenty additional online surveys that were administered to MBA students during the beginning of the Spring 2011

semester. The university Institutional Review Board approved the administration of the second set of surveys. In all cases, each survey had different comments and was unique. Under each comment, 21 subtopics were listed and each participant was asked to select the subtopic to which he/she thought the comment belonged to. In classifying these comments, the participants selected the best possible subcategory for each of the comments. Each survey was evaluated from two to five participants who acted as managers.

6.7.1 Analysis of Second Set of Surveys

For the second set of 20 human categorization surveys, the inter-reliability values were obtained: The results are listed below:

Table 6.9

Second Survey Set Inter-Rater Reliabilities

Survey	Raters	Kappa	ACI
V17	6	0.769254	0.770582
v18	3	0.64369	0.644482
v19	3	0.384082	0.389081
V20	3	0.699364	0.700743
V21	4	0.858146	0.858679
V22	3	0.821772	0.822666
V23	4	0.835845	0.836124
v24	3	0.776609	0.777836
v25	3	0.486352	0.489015
v26	4	0.816418	0.817114
v27	3	0.899906	0.900242
v28	6	0.76382	0.765033
v29	3	0.832358	0.833382
v30	3	0.727136	0.72781
v31	5	0.816406	0.817114
v32	3	0.788316	0.789417
v33	5	0.702961	0.704056
v34	5	0.697546	0.699087
v35	4	0.785131	0.786665
v36	3	0.799209	0.800039

Based on these results, survey number 19 and number 25 were discarded as their inter-rater reliability measures showed less than a substantial strength of agreement. The remaining surveys were used for the analysis of this research and the inter-rater reliability values are given below for these surveys.

Based on the inter-reliability rating values obtained of all 36 surveys, four surveys (#2,#10,#19, and #25) were deleted and not used for the classification experiment. Upon analysis of the first data set, it was determined that humans had difficulty categorizing the first subtopic: S01= Food Condition: Pizza is not very tasty. This subtopic stood out as an outlier compared to the other subtopics. As a result of this, comments related to this subtopic were also deleted from the classification experiment.

6.7.2 Human Categorization Results

After deleting four surveys, the surveys completed by 109 participants were analyzed. The results for the twenty subtopics indicate the following categorization measurements:

Table 6.10

Human Classification Subtopic Classification Scores

Subtopic	MacroF1 Score
S02	0.898
S03	0.854
S04	0.799
S05	0.779
S06	0.726
S07	0.463
S08	0.470
S09	0.556
S10	0.897
S11	0.927
S12	0.894

(table continues)

Table 6.10 (continued)

S13	0.953
S14	0.956
S15	0.815
S16	0.502
S17	0.577
S18	0.603
S19	0.846
S20	0.876
S21	0.900
Average Macro-F ₁	0.765
Micro-F ₁ Score	0.769
Accuracy	0.764

The analysis indicate that the top three subtopics that humans were able to categorize the most are S11, S13, 14 and the subtopics that humans were able to categorize the least are subtopics S07, S08, S16.

Table 6.11

Unique Human Classification Subtopics

Easier to Categorize Subtopics

S11=Facility: Employee uniforms are dirty.

S13=Facility: Environment is just too plain, too undecorated.

S14=Facility: Bathrooms are not clean.

Difficult to Categorize Subtopics

S07= Food Processing Methods: Pizza was not safe to eat (possible contamination).

S08= Food Processing Methods: Pizza does not look freshly made (cold pizza)

S16=Customer Service: When I need help, it takes too long for the server to notice.

Analyzing the main categories, the categorization scores improved. This indicated that humans could categorize better at a higher level of abstraction. The results are shown below:

Table 6.12

Main Categories Human Classification Scores

Main Category	MacroF1 Score
M1	0.913
M2	0.881
M3	0.970
M4	0.970
Average MacroF ₁ Score	0.934
MicroF ₁ Score	0.940
	0.938

The results indicate that humans could classify easily facility and customer service comments but experienced more difficult categorizing food processing comments.

6.8 LDA Categorization Procedure

The latent dirichlet allocation method is a text data mining method that is an extension of the pLSA method. The main difference between the two methods is that key mixture probabilities now follow the dirichlet multinomial distribution which is given as:

$$Dir(\alpha_1, \dots, \alpha_T) = \frac{\Gamma \sum_j \alpha_j}{\prod_j \Gamma \alpha_j} \prod_{j=1}^T P_j^{\alpha_j - 1} \quad (46)$$

As with pLSA, the latent dirichlet allocation method is based on the probabilistic topic model, the aspect model. Here, the probability of finding a word in a document is given by:

$$P(w_i) = \sum_{j=1}^T P(w_i/z_i = j)P(z_i = j) \quad (47)$$

where $P(Z_{i=j})$ =Probability that the topic= j for the i th word and $P(W_i/Z_{i=j})$ =conditional probability that of the word w_i given the fact that topic j was selected for the i th word.

In this study, the Gibbs sampling algorithm developed by Steyvers and Griffiths (2004) was used to perform LDA. The Gibbs sampling approach is a Markov chain Monte Carlo Procedure. Hogg, et al. (2005), explains the procedure for the Gibbs sampling. To start out with, a stream of X_o values are initialized at time $t=t_o$. Then a conditioned random variable Y_i/X_i is generated from a conditional distribution $f_{y/x}(y/x)$. This conditioned Y_i values are then substituted into another conditional distribution $f_{x/y}(x/Y)$ to generate a new set of conditioned X_i values or $X_i/Y_i \sim f_{x/y}(x/Y)$ and the process repeats itself. What is noted here is that the new state of the system only depends on the previous state and not on the past history and the movement from one state to another occurs on a random basis. These two concepts are the basis for this Gibbs sampling procedure being called a Markov chain Monte Carlo procedure—the future state only depends on the present state and not on its history, and moving from one state to another state occurs on a random basis.

In this Gibbs sampling approach, Steyvers and Griffins (2004) decided to define the posterior distribution of $P(z/w)$ as being proportional to the product of an estimate of $\varphi^{(j)} = P(w_i/z=j)$ and to an estimate of $\theta^{(d)} = P(z=j)$ In Griffiths and Steyvers's *Probabilistic Models*(2007) also defined the equation to select the next topic for a word token w_i given previous topics have been chosen for previous words as being *proportional* to these different terms and this equation is given below.

$$P(z_i = j | z_{-i}, w_i, d_i) \propto \frac{C_{w,j}^{WT} + \beta}{\sum_{w=1}^W C_{w,j}^{WT} + W\beta} \frac{C_{d,j}^{DT} + \alpha}{\sum_{d=1}^T C_{d,j}^{DT} + T\alpha} \quad (48)$$

The first term represents an estimate of $\varphi^{(j)} = P(w_i/z=j)$ and the second term represents an estimate of $\theta^{(d)} = P(z=j)$. They suggested that a good value for $\alpha = 50/T$, where T =number of

topics and $\beta=0.01$. In this particular research, these values were used to run every LDA trial. In this formula, C^{WT} and C^{DT} are matrixes with dimensions W by T and D by T respectively. C^{WT}_{wj} contains the number of times word w is assigned to topic j , not including the current instant i and C^{DT}_{dj} contains the number of times topic j is assigned to some work token in document d , not including the current instance i . The distributions $P(w_i/z=j)$ and $P(z=j)$ can be obtained from the C^{WT} and C^{DT} matrixes, either after each C^{WT} and C^{DT} is being computed or after a specific number of iterations have updated the C^{WT} and C^{DT} matrixes. The estimate probabilities for $\varphi^{(j)} = P(w_i/z=j)$ and $\theta^{(d)} = P(z=j)$ used to compute the probability of finding a word w_i are given by:

$$\Phi_i^{(j)} = \frac{C_{w,j}^{WT} + \beta}{\sum_{w=1}^W C_{w,j}^{WT} + W\beta} \quad (49)$$

$$\theta_j^{(d)} = \frac{C_{d,j}^{DT} + \alpha}{\sum_{j=1}^T C_{d,j}^{DT} + T\alpha} \quad (50)$$

The algorithm for LDA is based on Steyvers and Griffith (2004)'s implementation of the Markov Chain Monte Carlo Gibbs Sampling theory and this algorithm is considered a good technique to obtaining LDA results. This algorithm was used to obtain LDA results. How the above equations follow the Gibbs Sampling procedure is briefly outlined here. At time $t=t_o$, the topics are selected randomly for a word w_i in a document d_j . From these initial values the C^{WT} and C^{DT} matrices are computed. These matrices are then updated to compute the terms $\varphi^{(j)} = P(w_i/z=j)$ and $\theta^{(d)} = P(z=j)$. Then these terms are used to compute the probability of selecting the next topic for w_i in a document d_j or $P(z_{i=j}/z_{-i}, w_i, d_i)$. Once a new topic has been updated for each word w_i , new matrices C^{WT} and C^{DT} are computed and the process keeps going

until it converges and there is not much difference in the matrices after a number of iterations. Appendix A lists the flowchart of how this algorithm works.

6.8.1 LDA Topic Verification

The first step in the analyses of LDA categorization was to verify the topics extracted by the LDA Algorithm. The verification of these topics was achieved through the extraction of the documents numbers associated with each topic. Below is a sample of the LDA output for the four main categories for an LDA run that consisted of 50,000 iterations, $\alpha = 50/T$, where T =number of topics and $\beta=0.01$. The topics were to be classified into any one of the main categories: M1, M2, M3, and M4. Also, associated with the extraction of these topics, are the documents that were extracted (Table 6.14).

Table 6.13

LDA Extracted Topics

TOPIC_1	0.24284	TOPIC_2	0.25637
not	0.14915	Restaur	0.08049
food	0.08196	Look	0.05251
eat	0.05798	Employe	0.04378
tabl	0.03383	All	0.03987
dough	0.027	Dirti	0.03581
cook	0.02621	Clean	0.03536
process	0.02414	Facil	0.03355
TOPIC_3	0.25043	TOPIC_4	0.25036
pizza	0.28832	Custom	0.0949
crust	0.03897	Order	0.08073
top	0.0268	Service	0.05562
condition	0.0231	Server	0.04853
chees	0.02018	Time	0.03605
hard	0.01987	Serv	0.02511
am	0.01941	Long	0.0245

Table 6.14

LDA Extracted Documents

TOPIC_1	0.24285	TOPIC_2	0.25636
C2282	0.00365	C1015	0.00376
C0150	0.00317	C1174	0.00346
C0254	0.00317	C0140	0.00316
C1631	0.00286	C2238	0.00316
C0008	0.0027	C1384	0.00301
C0255	0.0027	C1173	0.00286
C1685	0.0027	C0258	0.0027
C0191	0.00254	C0764	0.0027
TOPIC_3	0.25043	TOPIC_4	0.25036
C0753	0.004	C0021	0.00369
C1932	0.00338	C1634	0.00323
C0045	0.00323	C0183	0.00308
C1163	0.00323	C0083	0.00292
C1165	0.00323	C1180	0.00292
C0108	0.00308	C0017	0.00277
C1128	0.00308	C0020	0.00262
C1164	0.00308	C0202	0.00262

These documents were then identified from the original list and were assigned to the topic number after considering all the possibilities.

Table 6.15

LDA Topic 1 Verification

Main Category	Subtopic	Document Number	Comment
M4	S18	C2282	I was very mad when I told the hostess I wanted a booth near the middle of the restaurant, and she sat me at a table at the back at the restaurant! I went to asked if we could be moved but she said that there was no booths open, when the next couple that came in was sat at a booth in the middle of the restaurant! I did not understand she did not give me the seating preference I wanted. The Customer Service was awful by ignoring my request and I will never return to a place where I will not get a simple seating request right.
M2	S07	C0150	Please be careful of dining at this pizza place. Not only was the pizza cold but it looked as if there was some kind of growth in the dough. I believe that it was definitely contaminated so I did not feel it was safe for I or my family to eat. They are definitely not keeping up with their food processing methods that would prevent this from happening and ensure more safe to eat pizzas.

(table continues)

Table 6.15 (continued)

M2	S09	C0254	I could not believe it when i saw it, the pizza i ordered actually had lumps in it. That is terrible and not good for business. Obviously, the pizza dough was not mixed well and that is a problem. The employees that prepared that should be fired. / / Major cause: Food processing methods / Specific cause: Pizza dough was not mixed well
M4	S18	C1631	I asked for a table because I have a leg injury--a broken ankle that's in a cast. It's much easier for me to get in/out of a chair than a booth. The waitress did not have a table available, when I could see empty tables over her shoulder. I had asked about those and she said they were reserved for call-ahead seating. The staff should be more accommodating for disabled folks like me and put me ahead of that list, especially when I sat there for 15 more minutes watching the same empty tables until the hostess finally seated me in one of them.
M2	S07	C0008	The past few times I came into your resteraunt I have received pizzas that have not been cooked properly but I am usually able to pop them in the oven for a few minutes and they are good to eat. This last time, though, I received a pizza with several hairs in it! This is not acceptable and leads to a possible contamination. I even saw the man who made the pizza not wash his hands, though I did not say anything at the time. I believe you need to seriously look into these issues as you are running risks for not only the customers but for your establishment.
M2	S07	C0255	The pizza I ordered was not safe to eat. I think it could have possibly contaminated my family and I!!! It was terrible. The worse pizza I have ever saw. / / Major cause: Food processing methods / Specific cause: Pizza was not safe to eat.
M2	S09	C1685	I ordered a cheese pizza today and the pizza dough had lumps in it so it was not mixed well. The food processing methods are not good because you are suppose to mix the dough probably. The pizza place needs to use a mixer to ensure the pizza dough is mixed correctly.
M2	S09	C0191	The pizza dough was not mixed well. I can still see spots of flower and lumps everywhere on this uncooked pizza. The cooks are not paying much attention to the dough preparation of the pizza and cooking it without it being mixed well. The food processing method needs to change for the better.

Table 6.16

LDA Topic 2 Verification

Main Category	Subtopic	Document Number	Comment:
M3	S14	C1015	I wish that I didn't have to go use the bathroom. The smell coming from it should have been a red stop sign for me already. As soon as I opened the door, I could see the water all over the floor. Paper towels were thrown everywhere except for the trash can. To make it even worst, there were puddles of water all by the toilet. When I was done and tried to flush the toilet, the toilet paper wouldn't even fully flush.
M3	S14	C1174	I cannot how unsanitary the bathroom were at this Pizza Restaurant. The toilets were not cleaned, there was only one stall with toilet paper, toilet paper was scattered throughout the restroom area, there was soap sprayed on the vanity tops, and wet paper towels in the sink. They need to keep a schedule of people continuously going into the restaurant to maintain its upkeep.

(table continues)

Table 6.16 (continued)

M3	S14	C0140	The bathrooms are unsanitary. The toilets have not been cleaned, the floors are covered in toilet paper and urine. The sinks have soap all over them and water is splashed all over the floors. The paper towel dispensers are empty and the trash is overflowing.
M3	S14	C2238	I am surprised to see how dirty and unsanitary the bathrooms are at the Pizza restaurant. The bathroom floors are disgusting, with toilet paper and paper towels thrown everywhere. The employees don't seem to care, I told them about it but there is no change. I believe the workers need to have a stronger work mentality in order to provide their customers with a cleaner bathroom environment. /
M3	S14	C1384	I visited the bathrooms at your establishment, and wow, what a mess. I was disgusted with the lack of cleanliness in the bathroom. I didn't even want to use the toilet because it was filthy. You were out of paper towels, water was on the floor, and it was a filthy mess. Your facility is lacking in one major area: clean bathrooms.
M3	S13	C1173	I cannot believe how dull and plain this Pizza Restaunt was. The walls were painted a plain white, there was no color to the walls, the cups and plates were even white, there were very few decorations and even the employees uniforms were dull. They need to incorporate some decorations and color into this place to make the environment more pleasant.
M3	S11	C0258	The employee that was helping me uniform was extremely dirty. I did not want her to touch my food at all. If her uniform is not clean no telling what else is not clean. Employees should look neat and clean when serving customers. A manager needs to address that please. / / Major Cause: Facility / Specific Cause: Employee's uniforms are dirty
M3	S14	C0764	When i went to the bathroom there was toilet paper every where and dirty toilets! I didnt want to use the bathroom because of how disgusting it was! I will complain to the manager and ask them how often they clean their bathrooms. There also was enough soap to wash my hands, I think they should pay more attention to the bathrooms and restock things and clean!

TABLE 6.17

LDA Topic 3 Verification

Main Category	Subtopic	Document Number	Comment
M1	S03	C0753	The pizza i ordered is a perfect size for alot of toppings. The pizza place didn't fill the pizza with enough toppings and for how much i am spending i think they should pile on the toppings. I will not tolerate such a bare pizza and demand my money back. This kind of pizza makes me mad i will not order from here ever again. There is hardely and pepperonies cheese or canadian bacon! This isn't a pizza! They shouldn't be so cheap in making a pizza!
M1	S02	C1932	The amount of money that is paid does not reflect the proper size of the pizza. The pizza should be made for to a bigger portion that will satisfy the customers hunger. The pizza maker shouldn't try to save money by diminshing the size of the pizza and amount of dough used. There needs to be a median to where the customer is satisfied with the amount of food and the pizza seller is happy with the supply of pizza being used.

(table continues)

Table 6.17 (continued)

M1	S03	C0045	The cheese on the pizza did not even cover the sauce. The pizza we ordered was a cheese pizza with pepperoni and mushrooms added and it was more like a crust covered with sauce and a few pepperonis and mushrooms. There were more places that you could see the sauce than places you saw cheese melted on top of the sauce. The condition of the pizza presentation was unappealing.
M1	S03	C1163	I cannot believe the quality of pizza topping at this Pizza restaurant. There is too much sauce, hardly any toppings, their isn't enough dough, they are scarce on the small pepperoni toppings, and the sauce seems to overpower the taste of anything else. They need a precise amount of measurement for each ingrediant used so that nothing is over-powering and there is enough to taste them all.
M1	S05	C1165	I cannot believe the crust quality at this Pizza Restaurant. The crust appears to be stale since the crust is too hard to even bite through. Once able to bite through this hard crust, it is very hard to chew. Either they have over-cooked the crust to extreme, or the dough is not fresh. They need to get the correct cooking temperature right and check on freshness of each patch of dough used for the crust.
M1	S03	C0108	The pizza toppings are not cut up but served on top of the pizza in much larger pieces. One slice of onion does not cover a pizza well. Lots of little piece cover the pizza so you have onion in all your bites, they are stingy with their ingredients.
M1	S03	C1128	Extra cheese means extra cheese!!!! I cannot believe you charge \$3.00 for extra cheese and I could still see the pizza sauce through the super thin layer of melted cheese. Do not even get me started on the hunt I had to go through looking for the sausage and onions. How dare you charge so much for your toppings when you put so little on your pizza?! I was outraged! We demanded a refund and left. We will not be back.
M1	S04	C1164	I cannot believe the taste of this sauce at the Pizza Restaurant. There are not a lot of ingrediants to the sauce that makes it too bland, it's soupy-like and makes the sauce seem like tomato juice. This sauce needs some type of spices, herbs or garlic to give it the flavor that the 'tomato juice' is lacking in. They should experiment in different sauce flavors with spices until they find the correct one(s).

Table 6.18

Topic 4 Verification

Main Category	Subtopic	Document Number	COMMENT
M4	S21	C0021	The last time I was at your resteraunt the server did not appropriately include the correct things in the bag. When I got home and opened the bag I found that my pizza was inside but there were no plates of napkins. I remember asking the cashier to put extra plates inside but to not include any is kind of stupid. This is the second time this has happened so I think that when you give your customer service training you should include a bit about making sure to not only ask how many plates a customer needs but to count the number of plates that way both the cashier and customer are sure they have been included.

(table continues)

Table 6.18 (continued)

M4	S19	C1634	The server took too long to attend to us. When I had to ask my friend "where's our waiter?" then one knows that the waiter is taking too long. Actually, it was about 30 minutes too long to get our food on a day that wasn't busy. I had looked over to the kitchen area, and noticed our food was ready, but it kept sitting there until the waiter delivered much later. The waiter took too long to notice our food was ready. My food was delivered cold, so I told her I didn't want. This was unacceptable.
M4	S20	C0183	The server brings the wrong order to a customer. The server must quickly fix the mistake whether it was the servers mistake or the customers. The customer service needs improvement between the server taking the customers order and checking the order before bringing it to the table.
M4	S20	C0083	The customer service in this place was terrible. The server brought the wrong order to our table,(she actually brought pasta and we ordered pizza), and when we told her it wasn't ours she got mad and said that the pasta was what we ordered! Then when she did bring out a pizza, we had ordered the meat lover's with Sicilian crust and she brought out a meat lover's with a Thin and Crispy crust, we were so hungry by this time that we accepted it, but when we again said it was wrong, she said "whatever".
M4	S20	C1180	I cannot believe how incompetant the servers were at this pizza Restaurant. The whole table was unable to get what they had ordered. They had mixed up my toppings with my husbands order of crust, non of our drinks we right and when they were corrected they were wrong again when we received a refill, my son got the wrong order altogether getting lasagna when he ordered meat lovers pizza, and I had extra toppings on my pizza that I didn't ask for. They servers need to make sure they write down each order and it isn't lost in translation when they go to the cooks to be sure they get customers food correct.
M4	S16	C0017	The last visit at your resteraunt I made did not go well at all. After we were served the server vanished for almost twenty minutes. I tried asking another server but he would not even fill my drink, as he was not my server. Finally afr 20 minutes the server came back and offered me the check. Next time your servers should provide better customer service or else I may take my business somewhere else.
M4	S20	C0020	The last time I was at your resteraunt I was actually given the wrong order. I got all the way home before realizing that my pizza had the wrong crust and had pineapples when I ordered Pepperoni. When I took the order back I was actually told that my order was mixed up with another gentleman's and I would have to wait either for the gentleman to come back or another pizza to be made! I am not sure how a mixup like this happens but I believe for next time you need to focus on Customer Service a bit more and maybe offer to have a customer whose order was messed up be put ahead of other customers since I had already been served and they had not.
M4	S17	C0202	Im allergic to garlic and I ask the server if a certain meal was made with any garlic . He was not sure and could not tell me any plate where he was sure it did not include garlic. I was not going to risk getting very ill so I decided not to order. The server did not seem to know what the meals included , he really didn't help me find a plate I could eat. He had poor customer service.

6.8.2 LDA Results

In this manner, the following LDA topics were identified to belong to the four categories M1, M2, M3, M4. The results for this trial are shown below:

Table 6.19:

<i>Topic Designation</i>	
Topic Designation-Trial 1	
Topic 1	M2
Topic 2	M3
Topic 3	M1
Topic 4	M4

To further support the designation a frequency distribution of the topics was also determined.

The frequency distribution for the above topics is shown in Table 6.20.

Table 6.20:

<i>Frequency Distribution of Main Categories</i>				
Frequency Distribution of Main Categories				
	Topic 1	Topic 2	Topic 3	Topic 4
M1	883	672	3223	640
M2	2280	755	1514	687
M3	1093	3846	620	597
M4	2032	1366	1128	4559

The frequency distribution serves to further verify the identification of the above topics. The results from the frequency distribution of topics for main categories yield similar results which are listed below:

Table 6.21:

<i>Topic Designation- Trial 1</i>	
Topic Designation-Trial 1	
Topic 1	M2
Topic 2	M3
Topic 3	M1
Topic 4	M4

Once the topics are identified, the resultant document topic distribution generated from the application of the LDA algorithm is analyzed. For each trial, the LDA algorithm was run for 50,000 iterations, with $\alpha=50/T$ and $\beta=0.01$. In LDA, one result is that a document has a distribution of topics associated with it. The topic with the highest frequency is selected as the topic predicted for that specific document. When a tie is shown, the document is assigned to the topic that appeared first. For every document, there's the actual category that should have been, the predicted category assigned by LDA, and the determination of whether the classification is correct. The classification results for the first trial are given below:

Table 6.22:

<i>LDA Main Categories Classification Scores</i>				
Trial 1				
Main Category	M1	M2	M3	M4
Precision	0.772	0.668	0.882	0.972
Recall	0.964	0.690	0.946	0.765
Macro-F1 Score	0.857	0.679	0.913	0.856
Average Macro-F1 Score	0.826			
Micro-F1 Score	0.835			
total correct	1603			
Accuracy	0.835			

The results above indicate that participants found it easier to categorize M3 comments and more difficult to classify M2 comments. The results also indicate that compared to humans, humans categorize better than the LDA algorithm for these categories.

The LDA algorithm was run for five trials, at 50,000 iterations per trial. The summary of main category classification by LDA is shown below.

Table 6.23

LDA Main Categories Macro-F1 Scores

MacroF1 Scores	M1	M2	M3	M4	Average
Trial 1	0.857	0.679	0.913	0.856	0.826
Trial 2	0.866	0.797	0.926	0.907	0.874
Trial 3	0.715	0.800	0.764	0.864	0.786
Trial 4	0.847	0.786	0.849	0.863	0.836
Trial 5	0.856	0.759	0.920	0.876	0.853
Average Macro-F1 Score	0.828	0.764	0.874	0.873	0.835

Table 6.24

LDA Main Categories Classification Scores

MicroF1 Score	Average Macro-F1 Score	Micro-F1 Score	Accuracy
Trial 1	0.826	0.835	0.835
Trial 2	0.874	0.881	0.881
Trial 3	0.786	0.793	0.793
Trial 4	0.836	0.841	0.841
Trial 5	0.853	0.859	0.859
Average	0.835	0.842	0.842

The results indicate that LDA has an average accuracy rate of 0.842 for the main categories which is less than the human accuracy rate of 0.938. The LDA results for the subtopics are summarized below:

Table 6.25:

LDA Subtopic Summary Classification Scores

	Total Correct	Total Predicted	Total Actual	Micro-F1	Accuracy	Average Macro-F1
Trial 1	1205	1920	1920	0.628	0.628	0.580
Trial 2	1193	1901	1920	0.624	0.621	0.560
Trial 3	1262	1901	1920	0.661	0.657	0.603
Trial 4	1184	1920	1920	0.617	0.615	0.550
Trial 5	1284	1920	1920	0.669	0.667	0.623
Average	1226	1912	1920	0.640	0.638	0.583

The results indicate that LDA has an average accuracy rate of 0.639 for the subtopics which is less than the human accuracy classification rate of 0.764. For a more detailed view of the performance classification of each subtopic, the following precision, recall and Macro-F1 scores were computed for each subtopic.

Table 6.26

LDA Subtopic Detailed Classification Scores

Precision	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11
trial 1	0.85	0.69	0.77	0.49	0.36	0.54	0.47	0.06	0	0.66
trial 2	0.86	0.47	0	0.73	0.16	0.63	0.56	0.53	0.43	0
trial 3	0.91	0.75	0.78	0.77	0	0.57	0.48	0.87	0.48	0.58
trial 4	0.27	0.63	0.71	0.79	0	0.66	0.77	0.63	0.24	0.59
trial 5	0.6	0.77	0.8	0.75	0	0.4	0.84	0.56	0	0.63
Average	0.7	0.66	0.61	0.71	0.1	0.56	0.62	0.53	0.23	0.49
Recall	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11
trial 1	0.87	0.78	0.81	0.82	0.38	0.43	0.52	0.01	0.09	0.84
trial 2	0.84	0.9	0	0.95	0.05	0.57	0.7	0.83	0.17	0
trial 3	0.88	0.8	0.79	0.94	0	0.61	0.64	0.86	0.15	0.79
trial 4	0.14	0.79	0.8	0.94	0	0.46	0.69	0.91	0.1	0.78
trial 5	0.91	0.76	0.83	0.93	0	0.45	0.71	0.86	0	0.86
Average	0.69	0.81	0.61	0.94	0.01	0.52	0.68	0.87	0.1	0.61

(table continues)

Table 6.26 (continued)

MacroF1-Score	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11
trial 1	0.86	0.74	0.79	0.62	0.37	0.48	0.49	0.02	0	0.74
trial 2	0.85	0.62	0	0.82	0.08	0.6	0.62	0.65	0.24	0
trial 3	0.89	0.78	0.79	0.85	0	0.59	0.55	0.87	0.22	0.67
trial 4	0.18	0.7	0.75	0.86	0	0.54	0.73	0.74	0.15	0.67
trial 5	0.72	0.76	0.82	0.83	0	0.42	0.77	0.68	0	0.73
Precision	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
trial 1	0.6	0.82	0.68	0.36	0.53	0.65	0.87	0	0.84	0.82
trial 2	0.42	0.86	0.67	0.72	0.5	0.79	0.86	0	0.7	0.75
trial 3	0.63	0.89	0.49	0	0.45	0.63	0.86	0	0.76	0.6
trial 4	0.65	0.53	0.47	0	0.49	0.6	0.81	0	0.79	0.66
trial 5	0.71	0.89	0.65	0.38	0.5	0.8	0.88	0.17	0.74	0.8
Average	0.6	0.8	0.59	0.29	0.49	0.69	0.85	0.03	0.77	0.72
Recall	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
trial 1	0.89	0.85	0.95	0.31	0.66	0.85	0.97	0	0.77	0.75
trial 2	0.89	0.83	0.93	0.74	0.71	0.86	0.95	0	0.73	0.78
trial 3	0.86	0.89	0.91	0	0.68	0.88	0.96	0	0.77	0.75
trial 4	0.85	0.79	0.91	0	0.76	0.86	0.98	0	0.79	0.78
trial 5	0.88	0.89	0.92	0.27	0.65	0.85	0.96	0.08	0.76	0.81
Average	0.87	0.85	0.91	0.25	0.7	0.86	0.96	0.02	0.76	0.78
MacroF1-Score	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
trial 1	0.72	0.84	0.8	0.33	0.59	0.74	0.92	0	0.8	0.78
trial 2	0.57	0.85	0.78	0.73	0.59	0.83	0.9	0	0.71	0.77
trial 3	0.73	0.89	0.63	0	0.54	0.73	0.91	0	0.77	0.66
trial 4	0.74	0.63	0.62	0	0.6	0.71	0.89	0	0.79	0.71
trial 5	0.78	0.89	0.76	0.32	0.56	0.83	0.92	0.11	0.75	0.8
Average	0.71	0.82	0.72	0.28	0.58	0.77	0.91	0.02	0.77	0.75

The results indicate that LDA has difficulty categorizing comment S06, S19 and S10, and found it easier to categorize comments S13, S17, S18, and S20. The description of these subtopics is listed in Table 6.27.

Table 6.27

Unique LDA Categorization Subtopics

LDA EASIER TO CATEGORIZE
S13=Facility: Environment is just too plain, too undecorated.
S17=Customer Service: Server is not helpful while I try to make my order decisions.
S18=Customer Service: Host is not willing to accommodate my seating preference.
S20=Customer Service: Server brings the wrong order (wrong crust, wrong toppings,
LDA DIFFICULT TO CATEGORIZE
S06= Food Processing Methods: Pizza is under or overcooked.
S19=Customer Service: My order takes too long to be served.
S10=Facility: The restaurant's entire waiting area looks dirty.

6.9 LSA RESULTS

The LSA results for the main categories are given in Table 6.26 and Table 6.27. The tables illustrate that LSA-IDF is a better classifier than LSA-TF.

Table 6.28

LSA-IDF Main Categories Classification Scores

LSA-IDF ANALYSIS OF MAIN CATEGORIES					
	M1	M2	M3	M4	Average
Precision	0.67	0.88	0.88	0.94	0.84
Recall	0.92	0.67	0.91	0.84	0.84
Macro-F1	0.77	0.76	0.89	0.89	0.83
Accuracy	0.92	0.67	0.91	0.84	0.84
Overall Accuracy	0.84				
Micro-F1	0.84				

Table 6.29

LSA-TF-Main Categories Classification Scores

LSA-TF ANALYSIS OF MAIN CATEGORIES					
	M1	M2	M3	M4	Average*
Precision	0.40	0.00	0.76	0.73	0.63
Recall	0.81	0.00	0.71	0.76	0.76
Macro-F1	0.54	0.00	0.73	0.74	0.67
Accuracy	0.81	0.00	0.71	0.76	0.76
Overall Accuracy	0.59				
Micro-F1 Score	0.59				

*excludes M2

For the main categories, the LSA-IDF version was better than the LSA-TF version for an overall accuracy of 0.84 compared to 0.59 for the LSA-TF version. For the LSA-TF version, the category M2 fails to load into a factor. Category M4 loaded into factor 2 and factor 4 and their classification results were combined.

Table 6.30

LSA Subtopic Summary Classification Scores

LSA RESULTS ON SUBTOPICS						
	Micro-F1 Score	Accuracy	Correct	Unclassified	Predicted	Average Macro-F1 Score
LSA-TF Version	0.471	0.381	731	739	1181	0.435
LSA-IDF Version	0.767	0.669	1343	339	1581	0.739

LSA-IDF Version performed slightly better than LSA with an accuracy rate of 0.67 compared to LDA accuracy rate of 0.639. In the LDA-IDF version, subtopic S10 did not load into the factors. Subtopic S06 loaded into two factors. The problem with the LSA-IDF Version is that 339 documents out of 1920 were left unclassified. The LSA-TF Version, however, proved to be the worst categorization algorithm with an accuracy rate of 0.381 and with 739 documents being left unclassified. In this version, none of the S04 subtopics were predicted to be correct.

To analyze the effect on the specific subtopics, the precision, recall and Macro-F1 Scores were obtained on the categories.

Table 6.31

LSA Specific Subtopic Classification Scores

BASIS	S02	S07	S03	S20	S05	S19	S15	S17	S09	S11	S08
LSA-TF	F20.1	F20.2	F20.3	F20.4	F20.5	F20.6	F20.7	F20.8	F20.9	F20.10	F20.11
Subtopic	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
Precision	0.40	0.78	0.00	0.86	0.13	0.53	0.60	0.92	0.13	0.66	0.79
BASIS	F20.1	F20.2	F20.3	F20.4	F20.5	F20.6	F20.7	F20.8	F20.9	F20.10	F20.11
LSA-IDF	S09	S11	S14	S19	S12	S05	S18	S17	S15	S20	S21
Subtopic	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
Precision	0.98	0.84	0.93	0.88	0.83	0.65	0.94	0.95	0.00	0.90	0.93
BASIS	S02	S07	S03	S20	S05	S19	S15	S17	S09	S11	S08
LSA-TF	F20.1	F20.2	F20.3	F20.4	F20.5	F20.6	F20.7	F20.8	F20.9	F20.10	F20.11
Subtopic	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
Recall	0.62	0.59	0.00	0.82	0.02	0.41	0.30	0.71	0.01	0.46	0.62
BASIS	F20.1	F20.2	F20.3	F20.4	F20.5	F20.6	F20.7	F20.8	F20.9	F20.10	F20.11
LSA-IDF	S09	S11	S14	S19	S12	S05	S18	S17	S15	S20	S21
Subtopic	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
Recall	0.86	0.76	0.77	0.93	0.42	0.44	0.68	0.82	0.00	0.82	0.91
BASIS	S02	S07	S03	S20	S05	S19	S15	S17	S09	S11	S08
LSA-TF	F20.1	F20.2	F20.3	F20.4	F20.5	F20.6	F20.7	F20.8	F20.9	F20.10	F20.11
Subtopic	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
Macro-F1	0.49	0.67	0.00	0.84	0.04	0.46	0.40	0.80	0.02	0.54	0.69
BASIS	F20.1	F20.2	F20.3	F20.4	F20.5	F20.6	F20.7	F20.8	F20.9	F20.10	F20.11
LSA-IDF	S09	S11	S14	S19	S12	S05	S18	S17	S15	S20	S21
Subtopic	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
Macro-F1	0.92	0.80	0.84	0.90	0.56	0.52	0.79	0.88	0.00	0.86	0.92
BASIS	S21	S12	S18	S14	S06	S04	S16	S13	S10	Average	
LSA-TF	F20.12	F20.13	F20.14	F20.15	F20.16	F20.17	F20.18	F20.19	F20.20		
Subtopic	S13	S14	S15	S16	S17	S18	S19	S20	S21		
Precision	0.95	0.92	0.15	0.03	0.55	0.93	0.17	0.66	0.94	0.53	

(table continues)

Table 6.31 (continued)

BASIS	F20.12	F20.13	F20.14	F20.15	F20.16	F20.17	F20.18	F20.19	F20.20	
LSA-IDF	S02	S13	S03	S07	S04	S08	S10	S06	S16	
Subtopic	S13	S14	S15	S16	S17	S18	S19	S20	S21	
Precision	0.99	0.78	0.94	0.31	0.75	0.92	0.65	0.95	0.94	0.80
BASIS	S21	S12	S18	S14	S06	S04	S16	S13	S10	Average
LSA-TF	F20.12	F20.13	F20.14	F20.15	F20.16	F20.17	F20.18	F20.19	F20.20	
Subtopic	S13	S14	S15	S16	S17	S18	S19	S20	S21	
Recall	0.20	0.64	0.07	0.01	0.48	0.57	0.09	0.54	0.46	0.41
BASIS	F20.12	F20.13	F20.14	F20.15	F20.16	F20.17	F20.18	F20.19	F20.20	
LSA-IDF	S02	S13	S03	S07	S04	S08	S10	S06	S16	
Subtopic	S13	S14	S15	S16	S17	S18	S19	S20	S21	
Recall	0.80	0.85	0.69	0.11	0.76	0.89	0.77	0.77	0.84	0.67
BASIS	S21	S12	S18	S14	S06	S04	S16	S13	S10	Average
LSA-TF	F20.12	F20.13	F20.14	F20.15	F20.16	F20.17	F20.18	F20.19	F20.20	
Subtopic	S13	S14	S15	S16	S17	S18	S19	S20	S21	
Macro-F1	0.33	0.75	0.10	0.02	0.51	0.71	0.12	0.59	0.62	0.45
BASIS	F20.12	F20.13	F20.14	F20.15	F20.16	F20.17	F20.18	F20.19	F20.20	
LSA-IDF	S02	S13	S03	S07	S04	S08	S10	S06	S16	
Subtopic	S13	S14	S15	S16	S17	S18	S19	S20	S21	
Macro-F1	0.89	0.82	0.80	0.17	0.76	0.90	0.70	0.85	0.89	0.73

The results on the subtopics indicate that LSA-TF version found it difficult to categorize S04, S06, S10, S13, S16, S15, and S19 comments. It found it easier to categorize, S05, S09 and S14 comments. The LSA-IDF version found it easier to categorize more comments than the LSA-TF version. It found it easier to categorize S02, S04, S05, S09, S12, S13, S15, S18, S20, and S21. It found it difficult to categorize S10 and S16 comments. The results are summarized in Table 6.30 and Table 6.31.

Table 6.32

Unique LSA-TF Categorization Subtopics

LSA -TF EASIER TO CATEGORIZE
<p>S05= Food Condition: Pizza crust is too hard. S09= Food Processing Methods: Pizza dough was not mixed well (has lumps inside). S14=Facility: Bathrooms are not clean.</p>
LSA-TF DIFFICULT TO CATEGORIZE
<p>S04= Food Condition: Pizza has either too much or too little amount of spices. S06= Food Processing Methods: Pizza is under or overcooked. S10=Facility: The restaurant's entire waiting area looks dirty. S13=Facility: Environment is just too plain, too undecorated. S16=Customer Service: When I need help, it takes too long for the server to notice. S15=Customer Service: Employees are impolite or rude. S19=Customer Service: My order takes too long to be served.</p>

Table 6.33

Unique LSA-IDF Categorization Subtopics

LSA-IDF EASIER TO CATEGORIZE
<p>S02= Food Condition: Pizza size is too small. S04= Food Condition: Pizza has either too much or too little amount of spices. S05= Food Condition: Pizza crust is too hard. S09= Food Processing Methods: Pizza dough was not mixed well (has lumps inside). S12=Facility: Kitchen area looks dirty, disorganized, messy. S13=Facility: Environment is just too plain, too undecorated. S15=Customer Service: Employees are impolite or rude. S18=Customer Service; Host is not willing to accommodate my seating preference. S20=Customer Service: Server brings the wrong order (wrong crust, wrong toppings, S21=Customer Service: Pizza is not served appropriately (plates, napkins, knives, etc.).</p>
LSA-IDF DIFFICULT TO CATEGORIZE
<p>S10=Facility: The restaurant's entire waiting area looks dirty. S16=Customer Service: When I need help, it takes too long for the server to notice.</p>

6.10 Discussion

Humans found it easier to classify documents under the S11, S13, and 14 subtopics and difficult to categorize documents under the S07, S08, and S16 subtopics. For LSA, the LSA-TF version found it difficult to categorize S04, S06, S10, S13, S16, S15, and S19 comments. It found it easier to categorize, S05, S09, and S14 comments. The LSA-IDF version found it difficult to categorize S10 and S16 documents. The LSA-IDF version found it easier to categorize more comments than the LSA-TF version. It found it easier to categorize S02, S04, S05, S09, S12, S13, S15, S18, S20, and S21. LDA found it difficult to categorize S06, S19 and S10 categories. It found it easier to categorize documents under the S13, S17, S18, and S20 categories. The results are summarized in Table 6.32. In this table, it can be seen clearly that LSA TF version finds it more difficult to classify more subtopics than any other method. Subtopic S10 was difficult to categorize for both LDA and LSA. Subtopic S16 was difficult to categorize for both humans and LSA. Subtopic S10 and S16 was difficult to categorize for LSA under any version. For unknown reasons, humans found it easier to categorize facility related comments than any other method. Three specific comments, S18, S20 and S13 were found to be easier to categorize by both LSA-IDF and LSA.

The results clearly indicate that at the specific level, there is no clear agreement between all the methods for classification. The algorithms and humans differ in the subtopics that they find difficult and easier to categorize. At the higher abstract level, only the LSA-TF method had difficulty categorizing main categories. As can be seen in Table 6.29, the main category M2: Food Processing had difficulty loading into a factor. For all the other methods, it was easier to classify the main categories than to classify the subtopics.

Table 6.34

Summary of Unique Categorization Subtopics for All Methods.

LDA EASIER TO CATEGORIZE	LDA DIFFICULT TO CATEGORIZE
<p>S13=Facility: Environment is just too plain, too undecorated.</p> <p>S17=Customer Service: Server is not helpful while I try to make my order decisions.</p> <p>S18=Customer Service: Host is not willing to accommodate my seating preference.</p> <p>S20=Customer Service: Server brings the wrong order (wrong crust, wrong toppings,</p>	<p>S06= Food Processing Methods: Pizza is under or overcooked.</p> <p>S19=Customer Service: My order takes too long to be served.</p> <p>S10=Facility: The restaurant's entire waiting area looks dirty.</p>
LSA -TF EASIER TO CATEGORIZE	LSA-TF DIFFICULT TO CATEGORIZE
<p>S05= Food Condition: Pizza crust is too hard.</p> <p>S09= Food Processing Methods: Pizza dough was not mixed well (has lumps inside).</p> <p>S14=Facility: Bathrooms are not clean.</p>	<p>S04= Food Condition: Pizza has either too much or too little amount of spices.</p> <p>S06= Food Processing Methods: Pizza is under or overcooked.</p> <p>S10=Facility: The restaurant's entire waiting area looks dirty.</p> <p>S13=Facility: Environment is just too plain, too undecorated.</p> <p>S16=Customer Service: When I need help, it takes too long for the server to notice.</p> <p>S15=Customer Service: Employees are impolite or rude.</p> <p>S19=Customer Service: My order takes too long to be served.</p>
LSA-IDF EASIER TO CATEGORIZE	LSA-IDF DIFFICULT TO CATEGORIZE
<p>S02= Food Condition: Pizza size is too small.</p> <p>S04= Food Condition: Pizza has either too much or too little amount of spices.</p> <p>S05= Food Condition: Pizza crust is too hard.</p> <p>S09= Food Processing Methods: Pizza dough was not mixed well (has lumps inside).</p> <p>S12=Facility: Kitchen area looks dirty, disorganized, messy.</p> <p>S13=Facility: Environment is just too plain, too undecorated.</p> <p>S15=Customer Service: Employees are impolite or rude.</p> <p>S18=Customer Service: Host is not willing to accommodate my seating preference.</p> <p>S20=Customer Service: Server brings the wrong order (wrong crust, wrong toppings,</p> <p>S21=Customer Service: Pizza is not served appropriately (plates, napkins, knives, etc.).</p>	<p>S10=Facility: The restaurant's entire waiting area looks dirty.</p> <p>S16=Customer Service: When I need help, it takes too long for the server to notice.</p>
HUMANS EASIER TO CATEGORIZE	HUMANS DIFFICULT TO CATEGORIZE
<p>S11=Facility: Employee uniforms are dirty.</p> <p>S13=Facility: Environment is just too plain, too undecorated.</p> <p>S14=Facility: Bathrooms are not clean.</p>	<p>S07= Food Processing Methods: Pizza was not safe to eat (possible contamination).</p> <p>S08= Food Processing Methods: Pizza does not look freshly made (cold pizza)</p> <p>S16=Customer Service: When I need help, it takes too long for the server to notice.</p>

The overall classification results are listed in Table 6.33. The accuracy rate for categorizing

documents for the subtopics are: .764 for Humans, .639 for LDA, 0.381 for LSA-TF version and .67 for LSA-IDF version. One important fact to notice is that humans are not perfect classifiers at any abstract level. Even though much research has focused on developing algorithms to simulate human performance, the potential exists for these algorithms to exceed human performance as it can be clearly seen that humans are not 100% perfect in classification performance. This may be attributed to the fact that not all humans can generate clear coherent comments and not all humans can classify perfectly generated comments. At the lower abstract level, humans outperform the LSA and LDA algorithms in classification. This is followed by LSA-IDF then by LDA. The LSA-TF version performs the worst in categorizing at the lower abstract level. At the higher abstract level, the accuracy rate for categorizing documents for the main categories are: 0.938 for Humans, 0.835 for LDA, 0.84 for LSA-IDF version and 0.59 for LSA-TF version. At the higher abstract level, LDA and LSA-IDF perform very similar, but the LSA-TF version performs the worst. The results indicate that humans outperform the computer algorithms in classification for any abstract level. The results also indicate that LSA-IDF version outperforms LDA, but the LSA-TF version is the worst classifier out of all the methods.

Table 6.35

Overall Summary of Results

Human Categorization on		Human Categorization on Main Categories	
Average Macro-F1 Score	0.77	Average MacroF1 Score	0.93
Micro-F1 Score	0.77	MicroF1 Score	0.94
Accuracy	0.76	Accuracy	0.94
LDA Categorization on Subtopics		LDA Categorization on Main Categories	
Average Macro-F1 Score	0.58	Average MacroF1 Score	0.84
Micro-F1 Score	0.64	MicroF1 Score	0.84
Accuracy	0.64	Accuracy	0.84

(table continues)

Table 6.35 (continued)

LSA-IDF Categorization on		LSA-IDF Categorization on Main Categories	
Average MacroF1 Score	0.74	Average MacroF1 Score	0.83
MicroF1 Score	0.77	MicroF1 Score	0.84
Accuracy	0.67	Accuracy	0.84
LSA-TF Categorization on Subtopic		LSA-TF Categorization on Main Categories	
Average MacroF1 Score	0.44	Average MacroF1 Score	0.67
MicroF1 Score	0.47	MicroF1 Score	0.59
Accuracy	0.38	Accuracy	0.59

The results indicate that very little support for the proposition presented in chapter 5 that states: Human Category Performance is closer to LSA performance than to LDA performance. At the lower abstract level, this proposition is not supported as LSA-IDF is closer to human performance than LDA. At the higher abstract level, this proposition has more support as both LDA and LSA-IDF perform similarly. The results simply cannot support the conclusion that humans categorize using a statistical approach (LDA) or a principal component approach (LSA).

The limitations for this research are as follows. In the human classification, the comments were generated in response to specific subtopics and main categories and for a specific setting. These comments were then categorized in a survey that only allowed only one option for categorization. These limitations were set to determine the accuracy of classification. In a more realistic business scenario, similar comments are generated at random and for a random number of unknown topics. Determining the accuracy of classification in this scenario would be difficult as the original comment may belong to more than one topic and there may also be many options available to classify a comment. Changing the context for customer generation could also result in different classification results. In the LDA algorithm,

categorizing a topic for a document is determined by randomly sampling one topic at a time for a word and selecting the topic with the highest probability as the topic to be assigned to that word. This process yields a frequency topic distribution for a document and the document is classified to be a particular topic by selecting the topic with the highest frequency. An alternative approach could be used by modifying the algorithm to sample two or more topics for a word. This process could yield a different frequency topic distribution for a document and different results. Also, two or more topics could be selected for a document. This is applicable as real comments sometimes concern more than one topic. Lastly, for LSA document loadings were forced as to yield one topic per document. This created a large number of documents that were not classified. A different document loading scheme could yield different results.

One of the limitations in this research is the input term-document matrix to the LSA and LDA algorithm. Transformation of this input term-document matrix may yield closer results to the classification performance of humans for both algorithms. For LSA, a quick analysis of the residuals between the original input term-document matrix and the filtered term-document matrix showed that the residuals did not follow a normal distribution. Transformation of the input term-document matrix may cause these residuals to follow the normal distribution. Also a unique feature of LSA is its orthogonality properties. Analyzing the LSA classification performance by using different matrix rotations could yield closer results to human classification performance for LSA.

The other limitation of this study is that participants were business students and MBA students who acted as managers when classifying the comments. Using MBA students as managers is relevant as several research studies have used MBA students as proxy for managers (Eylon, D., and Au. K. Y., 1999; Abramson, et al., 1996; Heuer et al., 1999; Pinkster, R., 2008).

However, this research could be more realistically duplicated by using real pizza customers and real pizza restaurant managers at the restaurant locations. Real pizza customers could generate comments through the first survey and real managers could classify these comments through the second survey and then these comments could be run through the LSA and LDA algorithms. It is unclear if the results would change dramatically.

CHAPTER 7

FUTURE RESEARCH DIRECTIONS

Researchers have recognized that some problems solved by human memory can be solved by automated information systems through the development of computer algorithms. Classification is one of those tasks that have expanded into the computer field through the development of many classifiers. These classifiers are being designed with the objective of being able to match and exceed human performance for processing large volumes of documents. The Internet is the perfect medium to apply these classifiers as it is a medium that handles massive amounts of documents on a daily basis that require classification for different purposes (e.g., block web pages, retrieve certain web pages, etc).

One of the areas our research has application potential is the retrieval of documents as performed by the Internet Google Search Engine. According to Griffiths et al., (2007), the Internet search engine may be a “more compelling metaphor for the human memory”. In the “Google and the Mind” article, Griffiths et al. (2007) explained that the human memory works similar to the PageRank algorithm of the Google search engine for the retrieval of information. When a query is performed on the human mind, the mind searches through a semantic network of words of interconnected pieces of information (i.e., words or concepts) and retrieves back the relevant pieces of information. Similarly, when a query is performed on the World Wide Web, the Internet search engines searches through a network of interconnected web pages and retrieves back only the web pages that are relevant.

The retrieval of web pages from the Internet is usually done via two steps: a) the identification of the web page that match a query and b) the ranking of web pages in terms of

their importance prior to displaying them to the user. The first step of this retrieval process is classification. Documents have to be classified as being relevant to the query or not. One area of future of research that could facilitate the retrieval of documents is the incorporation of LDA or LSA into the identification-of-web-page step in the PageRank algorithm. LDA and LSA are two options that could be incorporated into the PageRank algorithm to determine whether these two algorithms will be able to identify web pages better than the existing code of PageRank that handles the first step of the retrieval process.

Once LDA or LSA algorithms have been incorporated into a PageRank algorithm, similar research as was performed by Griffiths et al. (2007) can be duplicated. Griffiths et al. (2007) compared the fluency of word retrieval by humans to the PageRank algorithm that operated on a semantically constructed network and discovered that PageRank outperformed humans in fluency of the retrieval process.

As document classifiers, LDA and LSA can be used to enhance parental control features associated with the Internet. In this manner, web sites that need to be limited to children will be blocked and prevented from being displayed to children. Similarly, spam e-mails also need to be blocked. Future research areas could be also to investigate how LDA and LSA can be incorporated into algorithms designed to block spam e-mails or undesirable web pages.

Another area of research for future investigation is expanding LDA to analyze more than just one corpus of documents. There are situations where a corpora of documents are linked via topics that could use LDA analysis. For example a corpus of documents on high school sports could consist of topics “education” and “sports” while a corpus of higher education issues could consists of “education” and “tuition”. These two sets are linked by the topic “education.” Analyzing corpora with linked topics is an issue that can be solved by an extension of the regular

LDA method, the Hierarchical Dirichlet Process (HDP). In the regular LDA method, the assumption are that the key mixture probability distributions follow Dirichlet distributions. Using Bayes' theorem and these probabilities, the probability of finding a word in a document $P(w/\alpha,\beta)$ is computed. In sampling from mixture to obtain these probabilities, the sampling is said to be drawn from a Dirichlet Process, $DP(\alpha,\beta)$. In the Hierarchical Dirichlet Process, this assumption is true for one corpus. However, this is extended to multiple corpora. Thus, each corpus has a topic mixture that follows *Dirichlet*(α) and each topic has a probability word mixture that follows *Dirichlet*(β) and sampling is said to be done from a Dirichlet Process, $DP(\alpha,\beta)$. Teh et al. (2006), extended this idea to connect these different corpus to each other by claiming that the parameters α and β are random variables that also require sampling from a distribution. In particular, the value, β , also known as a base probability measure, was to be sampled from a different Dirichlet Process, $DP(\gamma,H)$. The value for the new base parameter, H , that Teh et al. (2006) used in an experiment was 0.5, γ was deemed to be a random variable \sim gamma(1,0.1), and α was a random variable \sim gamma(1,0.1). This hierarchical sampling idea can be further extended into having the new base parameter value H be sampled from another Dirichlet Process having other different parameters.

Since intensive research has been performed in attempting to model the cognitive process of how the human mind performs categorization tasks, many models of human categorization have emerged. These models “vary in terms of design, stimuli representation, attentional mechanism and decision processes” (Willis et al., 2006). All of these models are network models that assume that the stimuli are directed linked to a label (a category). In these networks, the inputs are usually features of stimuli and the outputs nodes represent the categories. Appendix A details some of the human categorization models that have evolved over time in the

field of human categorization. Further future research could be performed to determine how the models of human categorization are related to the classification performance of LDA and LSA algorithms. The simulation of the human categorization models could be compared to the performance of the LDA and LSA algorithms.

One thing to notice when comparing human categorization models to LDA and LSA algorithms is the input variables into the human categorization models. In most of the human categorization network models, the input consists of cues (or features) that exist in a stimulus that would be used by the network to categorize that stimulus. These cues are usually visual and from physical objects. But in documents classification, cues are not that readily visible in a document. However, cues do exist in a document and they need to be extracted via a computer algorithm prior to being processed in a human categorization network for classification purposes. To compare the performance of document classification via LSA or LDA against a typical human categorization model, the extraction of cues steps need to be performed first prior to processing these cues through the human categorization network model. LDA and LSA are algorithms that already perform topic extraction where words are selected to represent topics. Future research could be to assess how these extracted words from LSA or LDA could serve as representative cues of a document that would allow this document to be identified via a human categorization model.

Another area of future research is to explore the similarity between the networks developed for human classification and the networks developed for data processing. A close inspection of some of the models reveals that there is not much difference between the two types of networks. As an illustration here, the Attention Learning Covering Map (ALCOVE) human categorization network model developed by Kruschke (1992) in an attempt to explain exemplar

theory of categorization is compared to the Artificial Neural Networks (ANN) that are used in classification of databases by information systems. The ALCOVE (attention learning covering map) model is a three layer hierarchical category learning model that relates the stimulus to an “exemplar” (denoted as a hidden node in the network) which is then used to select the category for that stimulus. In this model, an exemplar is activated based on the psychological distance, as measured by Nosofsky (1986) similarity function from the generalized context model (GMC), between the exemplar and the stimulus node. Figure 1 shows the structure for the ALCOVE model where certain network parameters need to be specified: fixed specificity c , probability mapping constant ϕ , association weight learning rate λ_w and the attention-learning rates λ_a .

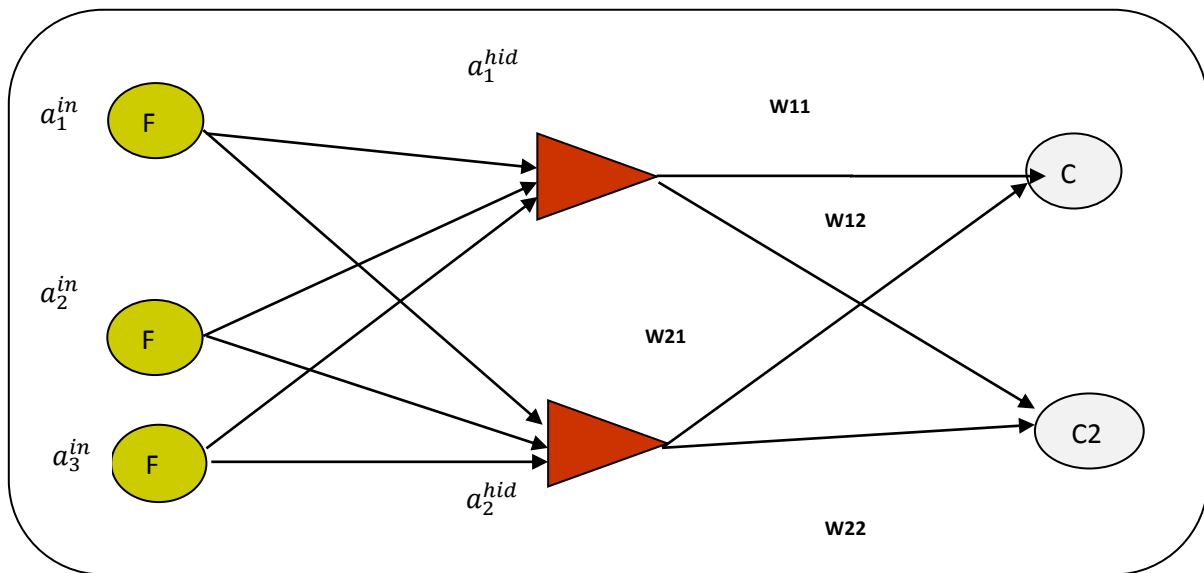


Figure 7.1 The attention learning covering map (ALCOVE) model.

The ALCOVE Model is very similar to an ANN in that its three hierarchical layers are similar to the three hierarchical layers of an ANN. In information systems, ANNs are used to solve many real world problems that involve classification, prediction, pattern recognition and non-linear problems that are difficult to solve through normal mathematical calculations.

(Almhdi et al., 2007) ANNs are used to evaluate databases in medical settings to classify patients, in business settings to predict profit or bankruptcy, etc. An ANN differs from the ALCOVE model in that there may be more hierarchical layers associated with an ANN where the output of one neuron becomes the input another neuron in the next layer. The ANN network also differs from the ALCOVE model in the type of transfer functions used to transform inputs to outputs. The figure below shows a basic three layer hierarchical structure of an ANN.

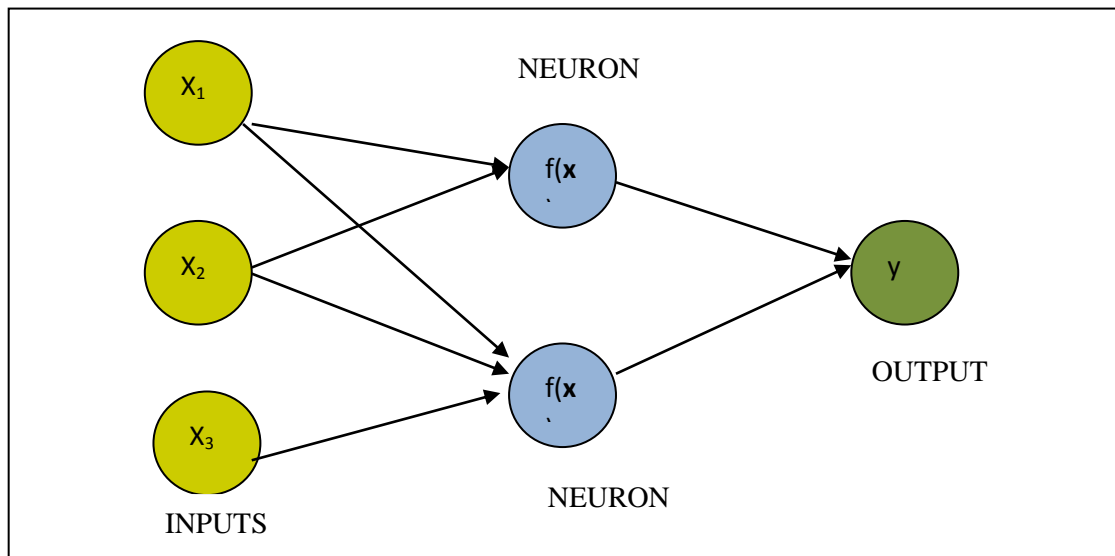


Figure 7.2 Artificial neural network.

Another difference between an ANN model and the ALCOVE model is that an ANN can be classified as a feed forward network or a recurrent network while the ALCOVE is only considered to be a feed forward network. In feed forward, signals flow from the input of to the output node in one direction and from layer to layer. In recurrent networks, the output of a neuron is sent back either to the same neuron or to a preceding neuron. In this way, ANNS can perform supervised learning and unsupervised learning. In supervised learning, the weights in the network are adjusted based on the output (i.e., the actual output is compared to an expected

output and its difference is used to adjust the weights of the network.). In unsupervised learning, the input determines the weights of the network.

In summary, both of these networks perform classification and are very similar in structure but differ in the types of internal functions used and their use as feedback or recurrent networks. Recognizing the similarity that exists within categorization networks in the two different academic fields is an area of research that needs to be explored. The field of cognitive science and information systems is highly linked and as we attempt to improve the processing of data in information systems, we cannot afford to ignore the development of networks that exist within the human categorization field.

One of the limitations in this research is form of the input term-document matrix to the LSA and LDA algorithm. Transformation of this input term-document matrix may yield closer results to the classification performance of humans for both algorithms. Also for LSA, matrix orthogonality plays a role on the results. Thus, transformation of the input term-document matrix for LSA and LDA and the effect of different matrix rotations for LSA are areas for further investigation.

In the LDA algorithm, sampling is performed to select a topic for a word and in the process, a topic distribution for a document is created. From this distribution, the topic with the highest frequency is assigned to the document. However, real comments sometimes concern more than one topic. A further area of investigation is to modify the algorithm as to sample two or more topics at a time by using joint probabilities and thus create a different topic distribution for a document. Furthermore, two or more topics at a time could be assigned to a document from this topic frequency distribution.

Lastly, further future research could be performed to understand the theoretical basis for LDA classification and retrieval of documents. Griffiths et al., (2007), Steyvers and Griffiths, (2008) show that the retrieval of document can incorporate Bayesian probabilities in the same manner that probabilistic topic models do. If we apply Bayesian probabilities to the retrieval of web pages by a search engine, the process can be explained in terms of Bayesian probabilities. Let w = web page that is relevant to query. Let q = data that represents query. Using Bayesian probabilities, the goal in information retrieval is to find the web pages with the highest probabilities $P(w/q)$ and list them.

$$P(w/q) = \frac{P\left(\frac{q}{w}\right)P(w)}{\sum_{w \in W} P\left(\frac{q}{w}\right)P(w)} \quad (51)$$

where w is the set of all web pages. Under this probability, $P(q/w)$ represents the first step of the retrieval process. $P(q/w)$ is the probability that a given web page will have data relevant to the query. If a web page does not have any information relevant to the query, then $P(q/w)$ is zero, otherwise $P(q/w)$ is constant for all web pages (Griffiths et al., 2007). But this issue of the Bayesian probability $P(q/w)$ being constant for all web pages (that contains the words in a query) also merits further investigation. If one web page contains two words that are part of the query and another similar word page contains ten words that are part of the query, how can you justify that they have the same probability $P(q/w)$? Analysis of the second step of the retrieval process also has some issues. $P(w)$ represents the second step of the retrieval process. $P(w)$ is a measure of the importance of that webpage. The PageRank algorithm measures importance by the number of links associated with a webpage. The problem is that some links are deemed more important than others in the interconnected world of web pages. For information retrieval, what is considered to be $P(w)$, a measure of importance, is an issue that merits further investigation.

This importance issue affects how the web pages are ranked prior to displaying them to the user. Thus, the application of Bayesian probabilities to the retrieval of information is an issue that merits further investigation.

APPENDIX A
MODELING OF HUMAN CATEGORIZATION

In the current literature of human categorization, various attempts have been made to model the cognitive process of the human mind involved in a categorization task. This Appendix is an overview of some of the simulation models that exist in the field of human categorization. It is presented here to provide the reader with an understanding of the human categorization models that may be related to information systems models.

According to Willis et al., (2006), “Models of categorization exist and they vary in terms of design, stimuli representation, attentional mechanism and decision processes.” All of them assume that the stimuli are directly linked to a label. Modeling the human cognitive categorization process has evolved in terms of networks that have been used in simulation studies since 1988. One of the first network models that was developed was the basic component-cue categorization model that was derived by Gluck and Bower (1988). This model is a network model where an input node is provided for every feature of a stimulus and the output nodes represent the categories. Stimuli are represented as points in a multidimensional psychological space. A stimulus has many dimensions (e.g., size, weight, etc.) with specific values assigned to each dimension. The model consists of N input feature nodes and C output category nodes. There is an input node per feature and output node per category. If a feature is present, the input node i has an activation a_i^{in} of 1 otherwise it has an activation of “0”. The k th output node is activated when the k th category is chosen. The activation for the output node k is given by a_k^{out} . All input nodes are connected to all output nodes. Every connection in the model has a weight $w_{i,k}$, where $i \in N$ and $k \in C$. that represents the degree of association between the feature and the category. Usually, after initializing the model with initial connections weights being set to zero, these connection weights are “learned” in an algorithmic fashion through

several iterations. The learned degree of association between the feature and the category increases the value of the weight until stability is reached.

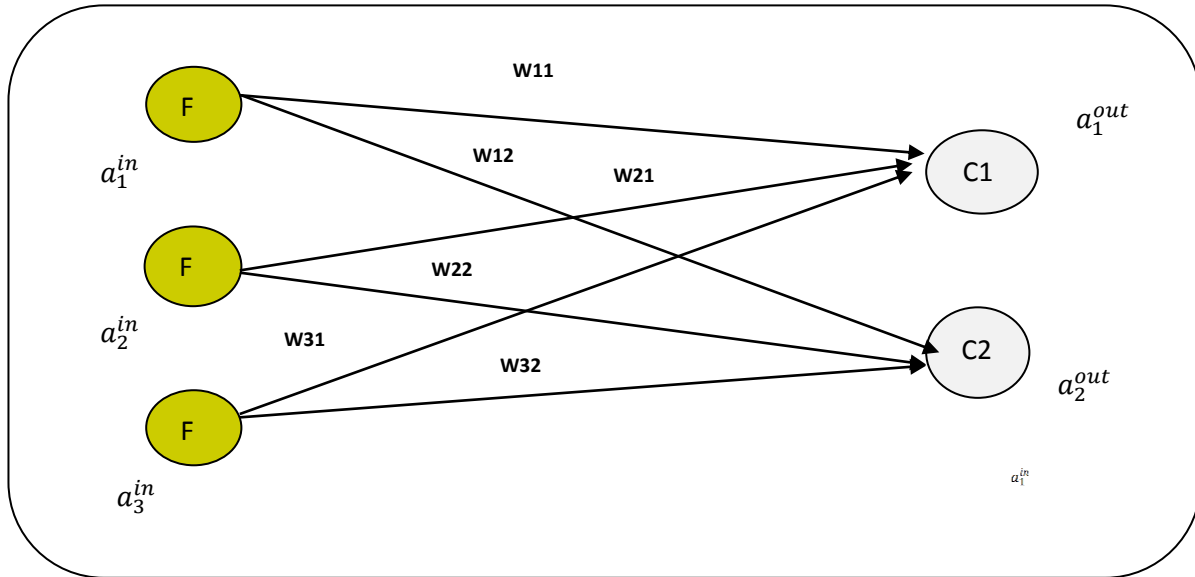


Figure A.1 The component-cue model.

A.1 The Attention Distinctive Input (ADIT) Model

The Attention to Distinctive Input (ADIT) Model is another network human categorization model that introduces a new shifting attention mechanism to the input stimuli nodes of the component-cue model. Kruschke (1994) explains the steps associated with this network and they are briefly mentioned here. For additional information, please see Kruschke (1994). Briefly, in the ADIT model, each input node i has an attention strength, α_i in addition to its activation value a_i^{in} . To begin, the model is activated by the presence of a stimulus with features. The attention strengths are normalized to account for the number of features. Each output node k is activated by the strength of the activation value a_k^{out} . After each output node k is activated, the network returns a feedback value t_k . The error for the network, E , is computed as a function of this t_k . After the error E is computed, attention strengths and connections

weights are adjusted. The network model incorporates base rates, prior category probabilities, in assigning to each category the probability of being selected. The model first computes an unbiased choice probability p_x for each category that is primarily a function of the output activation values. The network learns base rates r_k for each category. These base rates are incorporated into another equation to compute the revised choice probability of selecting a category k (pr_k). In the network the base rates r_k are initialized to $1/N$ and then they are incremented periodically.

Graphically, the ADIT model is shown below.

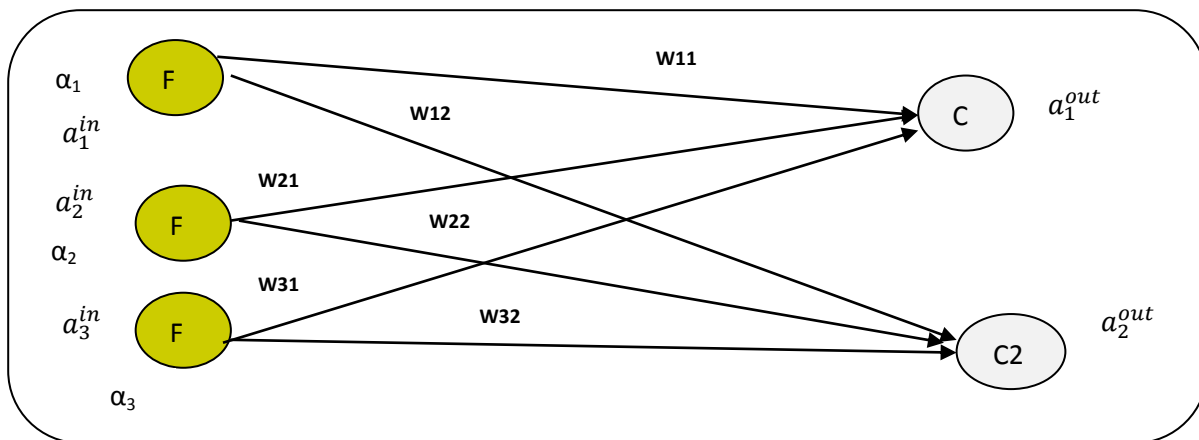


Figure A.2. ADIT model.

A.2 The Attention Learning Covering Map (ALCOVE) Model

The ALCOVE (attention learning covering map) model is a three layer hierarchical category human learning model that relates the stimulus to an “exemplar” (denoted as a hidden node in the network) which is then used to select the category for that stimulus. This network model was developed in an attempt to explain the exemplar theory of categorization. In this model, an exemplar is activated based on the psychological distance between the exemplar and the stimulus node. The model incorporates the Nosofsky (1986) similarity function from the generalized context model (GMC) to measure this psychological distance.

Kruschke (1992) outlines the procedure for the ALCOVE model. For detailed information on the procedure, please see Kruschke (1992). Here the activation of the hidden nodes are activated by the following equation that was also used in the GMC model and in the network models of Bluck and Bower (1988a, 1988b)

$$\alpha_j^{hid} = \exp[-c(\sum_i \alpha_i | h_{ji} - \alpha_i^{in} |^r)^{\frac{q}{r}}] \quad (A8)$$

Let position of jth hidden node be (hj1, hj2,) and let the activation of jth hidden node by a_j^{hid}
 c= specificity of the node,

r=constant for psychological distance

q= constant for similarity gradient

α_i = the attention strength of stimuli i .

α_i^{in} = activation of node i

In the ALCOVE model, the hidden nodes have diamond shape contours that is described by the similarity function. Each hidden node is connected to output nodes, response categories, with connections that have different weight values w_{ki} = weight from exemplar node k to category i .

A output node is activated by a formula from a hidden node. To compare model performance with human performance, these activation categories must be mapped into probabilities

As in the ADIT model, the dimensional attention strengths α and the association weights w_{ki} between exemplars and categories are learned in the ALCOVE model. Likewise, after each trial feedback is given as to the correct response and this is the t_k for a category code and an error rate E is computed as function of t_k . In the ALCOVE model, the association strengths after the exemplar and the attention strengths of the stimuli are changed after every trial. An illustration of the ALCOVE model is shown below:

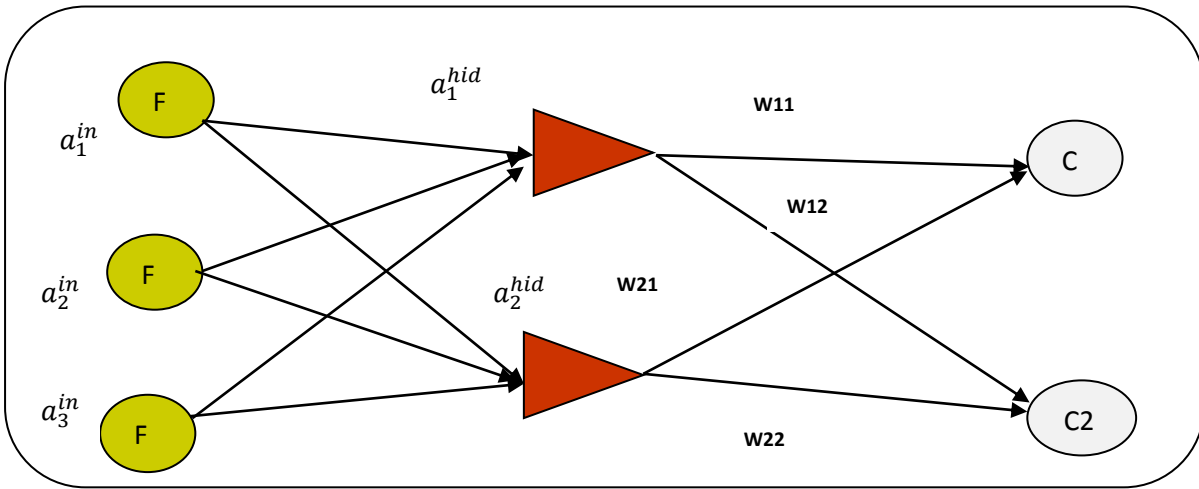


Figure A.3: The attention learning covering map (ALCOVE) model.

A.3 Casual Modeling

In attempting to explain human categorization, human categorization models have embraced casual theory to explain the effects of casual (prior) knowledge on categorization. Rehder (2003) explains that casual model theory does not only includes knowledge of features on an object or entity but also on knowledge of casual mechanisms that *link* those features. He gives an example of casual model theory where he explains that birds have wings, can fly, and can build nests up in trees. These features are linked in the following manner. Birds can fly *because* they have wings. Birds can build nests up in the tree *because* they can fly. The research questions that Redher (2003) attempted to explain are how much weight does each one of these linked features have in establishing category membership? What is the effect of combination of these linked features on categorization?

In explaining casual theory, features are said to belong to a casual chain model as illustrated in Figure 4. “c” is the probability that feature F1 exists. “b” is the probability that

feature F2 exists regardless of the previous feature. (This is if the variable “b” is found under feature F2.) “m” is the probability that one feature will cause the next feature

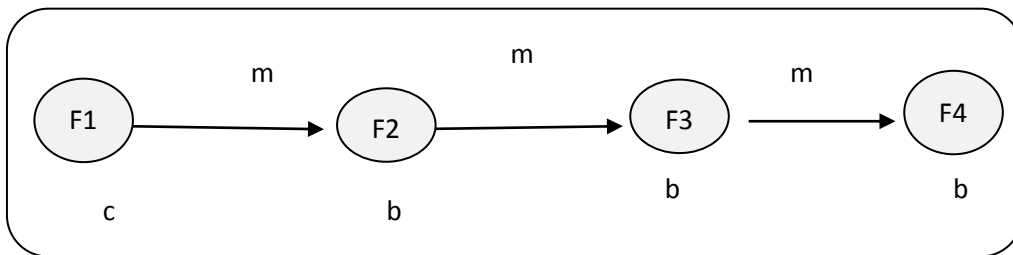


Figure A.4. Casual modeling of features in a stimulus.

Ahm et al. (2000) research tested the effect of casual relationships on categorizations. Here participants ranked three features X, Y, and Z that were linked in the following manner. X causes Y which causes Z. This research showed that participants rated worst the category when Y was missing than when Z was missing. When X was missing, it was rated worst of all. This indicated that in the casual relationship, the X feature was weighted more than Y feature than the Z feature.

Rehder’s (2003) research on casual modeling confirmed that some features are more salient than others and have an effect on categorization and that certain combination of features have an bigger effect on categorization. His research supported the claims that “A central claim of casual-model theory is that people’s knowledge of categories includes not just category features but also the representation of casual mechanisms that link those features” and “categorizers make classification decisions by estimating how likely an exemplar is to have been generated by a category’s casual model”(Rehder,2003).

A.4 Competition between Verbal and Implicit Systems (COVIS) Model.

The model of human categorization learning that attempts to explain the categorization process through the neuropsychological system is the Competition between Verbal and Implicit

Systems (COVIS) model. Before describing this system, a brief review of the important structures of the human brain is needed.

A.4.1 Brain Structures

The brain can be considered to be a body divided into different functional area. According to Ashby (2001), three areas of the human brain are of concern for categorization purposes: the frontal cortex, the basal ganglia and the medial temporal lobe. The frontal cortex comprises about a third of the entire brain and is connected to the sensory cortex which is associated with all sensory functions. For vision, the sensory cortex includes other brain areas called occipital cortex and much of the temporal and parietal cortex. For clearer illustrations of these brain areas, please visit the following websites to get images of the human brain:

<http://www.hss.iitb.ac.in/courses/HS435/Neurolinguistics.htm>

<http://inside-the-brain.com/2011/02/09/pioneering-brain-surgery-to-treat-tourettes/>

Within the frontal cortex, two structures, prefrontal cortex and the anterior cingulate, are considered to be critical for short term working memory and for executive attention, items that are considered important in category learning (Ashby, 2001). The frontal cortex is considered the locus of human reasoning and it has been shown to participate in category learning through explicit reasoning. Damage to the frontal cortex can be assessed through a special neuropsychological categorization test, the Wisconsin Card Sorting Test (WCST), and through neuroimaging studies. The basal ganglia are subcortical structures that exist within the brain. Studies have shown that people with diseases to the basal ganglia (e.g. Parkinson's disease, Huntington's disease) are impaired in category learning. (Knowlton et al. 1996a, 1996b) These structures are thought to be critical for procedural learning and memory. The medial temporal lobe consists of structures that are considered important in the memory of facts and events. It is

general believed that people use this recollection of events and facts in categorization of new stimuli. Amnesiacs usually experience damage to the medial temporal lobe. Studies have not shown concretely how exactly damage to the medial temporal lobe affects categorization learning.

According to Ashby and Maddox (2005), recent neuroscience evidence suggest that human category learning may be mediated by many different systems and not just one. Ashby and Maddox (2005) explain that evidence exists that shows that many neural structures participate in category learning and evidence also exist of multiple memory systems existing in the brain. Because category learning is believed to rely on memory, this evidence suggests that there are many different category learning systems. For once, the explicit reasoning process used in category learning that usually involves hypothesis testing, theory construction and testing of theories, may be system that is highly believed to be mediated frontal cortical structures. The implicit learning process may be mediated by another system.

In conclusion, many models have been developed throughout time in the field of human categorization. All these models have tried to model the actual classification process of the human brain. Ongoing research continues to determine how the actual mental process of how classification is performed. But the usefulness of the developed models is that many of these models can be linked to actual data classification models that are part of the information system field. The human categorization field seeks to provide a better understanding of the mental classification processes involved in the human brain. With this understanding, the development of better information systems that mimic the human mental processes can be achieved.

APPENDIX B
ONLINE SURVEY FOR PART I OF RESEARCH

A CONCEPTUALIZATION-ARTICULATION STUDY FOR QUALITY PROBLEMS

Part I: SURVEY INSTRUCTIONS

Consider the following made-up part of a cause-and-effect (Ishikawa, or fishbone) diagram. (NOTE: This example is presented to you for illustration purposes only; it does not correspond to a true cause-and-effect analysis. However, the task that you will be given in the main study corresponds to a realistic cause-and-effect analysis.)

An Illustration example

Major Cause Category Emotions (remember, this is made-up)

Specific Cause: Fits of anger (this is also made-up)

Customer Complaint Example 1: There is so much anger in the Pizza restaurant, people are ready to explode. They cannot put up with each other anymore and they start fighting. They should not get so emotional.

Customer Complaint Example 2: I cannot believe how angry everybody seems to get these days at the Pizza restaurant. They are yelling at each other, they are throwing pizza boxes. They are breaking things. We need to learn how to control our emotions.

Analysis of this example

- Description (articulation) statements include a problem symptom description, i.e., an example of everyday occurrence of the problem at the workplace (*they are yelling at each other, they start arguing*). Problem symptom description correspond to *anticipated* customer complaint comments.
- They clearly mention the specific cause (*anger, angry*).
- They clearly mention the related major cause category (*emotions, emotional*).

Your next step

You will now be asked to come up with your own description statements that are similar to Customer complaint Example 1 or Customer complaint Example 2. For each one of them, please make sure you:

- (1) describe a situation at the workplace that corresponds to a symptom related to the specific cause you are working on, as it would be articulated by a complaining customer.
- (2) mention the specific cause in your description, and

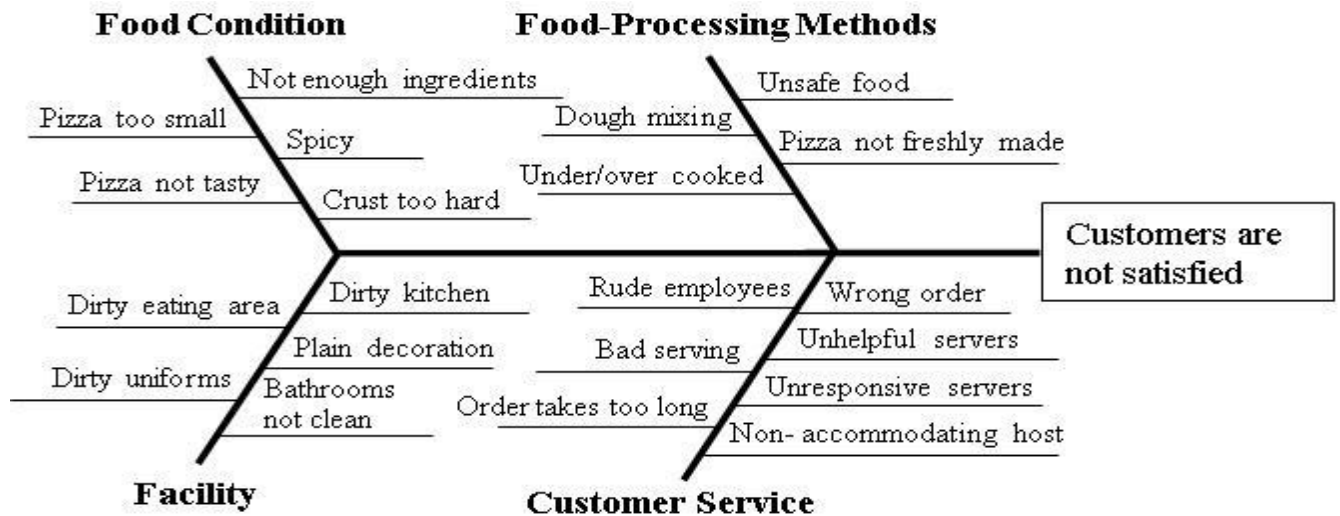
(3) mention the major cause category in your description.

How to earn your extra credit points

If you are a DSCI 2710 student in the Fall 2010 term, there is an opportunity for you to earn 10 points of extra credit (on the 800-point grading scale applicable to your DSCI 2710 course). By simply participating in this study you may not earn the available extra credit. Extra credit will be awarded as a bonus to participants whose input is of acceptable quality. Simply entering random text, irrelevant text, or submitting blank answers will not earn you the extra credit. Your input will be considered to be acceptable if it clearly addresses the three points listed in the previous section (i.e., [1] if it includes a relevant description of the situation, [2] if it mentions the corresponding specific cause, and [3] if it mentions the corresponding major cause category.) Input relevance will be measured by performing search engine-style (e.g., Google) automated queries. Our relevance thresholds will be relatively forgiving, so simply making an honest effort to produce meaningful input should be enough to earn you the extra credit points. If you are a DSCI 4520/5420 student, you will be awarded a 1% extra credit bonus for relevant participation following similar procedures.

PART 2: MAIN SURVEY

The survey below deals with quality issues related to customer satisfaction at a **sit-down pizza restaurant**. In this survey, quality is being defined as "meeting or exceeding customer's expectations." You are asked to provide description statements for each of the problem causes. Your description should include: (1) an everyday example of the problem cause as it would be worded by a complaining customer, (2) the specific cause, and (3) the major cause category.



Major Cause Category 1: **Food Condition**

Specific Cause 1.1 **Pizza is not very tasty.**

Your Complaint Description:

(Please type your description in the box below. Make sure you also mention the specific cause--taste, etc.-- and also the major cause category--food condition.)

[Show me an example](#))

Major Cause Category 1: **Food Condition**

Specific Cause 1.2 **Pizza size is too small.**

Your Complaint Description

(Please type your description in the box below. Make sure you also mention the specific cause--pizza size etc. --and also the major cause category--food condition.)

Major Cause Category 1: **Food Condition**

Specific Cause 1.3 **Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).**

Your Complaint Description

(Please type your description in the box below. Make sure you also mention the specific cause--not enough topping material etc.--and also the major cause category--food condition.)

Major Cause Category 1: **Food Condition**

Specific Cause 1.4 **Pizza has either too much or too little amount of spices.**

Your Complaint Description

(Please type your description in the box below. Make sure you also mention the specific cause --too spicy, etc.--and also the major cause category--food condition.)

Major Cause Category 1: **Food Condition**

Specific Cause 1.5 **Pizza crust is too hard.**

Your Complaint Description

(Please type your description in the box below. Make sure you also mention the specific cause--hard crust, etc.--and also the major cause category--food condition.)

Major Cause Category 2: **Food Processing Methods**

Specific Cause 2.1 **Pizza is under or overcooked.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

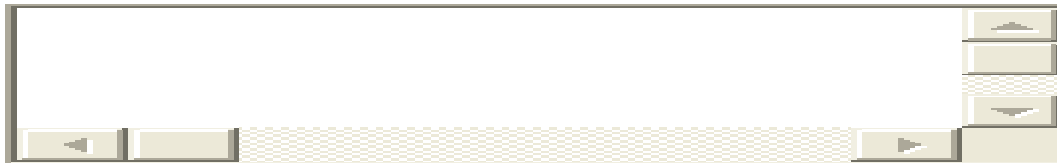
[\(Show me an example\)](#)

An empty complaint form template with a header area, a main text area, and a footer area with navigation buttons.

Major Cause Category 2: **Food Processing Methods**

Specific Cause 2.2 **Pizza dough was not mixed well (has lumps inside).**

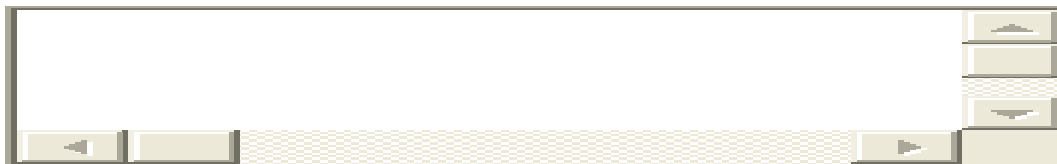
Your Complaint Description
(Do not forget to mention the specific cause and the major cause category!)

A complaint form template with the text "Major Cause Category 2: Food Processing Methods" and "Specific Cause 2.2 Pizza dough was not mixed well (has lumps inside)." in the header area.

Major Cause Category 2: **Food Processing Methods**

Specific Cause 2.3 **Pizza was not safe to eat (possible contamination).**

Your Complaint Description
(Do not forget to mention the specific cause and the major cause category!)

A complaint form template with the text "Major Cause Category 2: Food Processing Methods" and "Specific Cause 2.3 Pizza was not safe to eat (possible contamination)." in the header area.

Major Cause Category 2: **Food Processing Methods**

Specific Cause 2.4 **Pizza does not look freshly made (cold pizza).**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

Major Cause Category 3: **Facility**

Specific Cause 3.1 **The restaurant's entire eating area looks dirty.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

[\(Show me an example\)](#)

Major Cause Category 3: **Facility**

Specific Cause 3.2 **Employee's uniforms are dirty.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

Major Cause Category 3: **Facility**

Specific Cause 3.3 **Kitchen area looks dirty, disorganized, messy.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

Major Cause Category 3: **Facility**

Specific Cause 3.4 **Environment is too plain, too undecorated.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

Major Cause Category 3: **Facility**

Specific Cause 3.5 **Bathrooms are not clean.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

Major Cause Category 4: **Customer Service**

Specific Cause 4.1 **Employees are impolite or rude.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

[\(Show me an example\)](#)

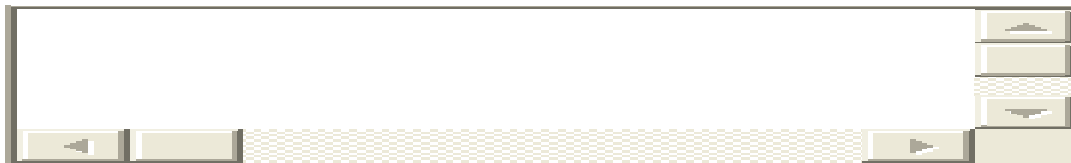


Major Cause Category 4: **Customer Service**

Specific Cause 4.2 **Host is not willing to accommodate my seating preference.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)



Major Cause Category 4: **Customer Service**

Specific Cause 4.3 **When I need help, it takes too long for the server to notice.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)



Major Cause Category 4: **Customer Service**

Specific Cause 4.4 **Server is not helpful while I try to make my order decisions.**

Your Complaint Description

(Do not forget to mention the specific cause and the major cause category!)

Major Cause Category 4: **Customer Service**

Specific Cause 4.5 **My order takes too long to be served.**

Your Complaint Description
(Do not forget to mention the specific cause and the major cause category!)

Major Cause Category 4: **Customer Service**

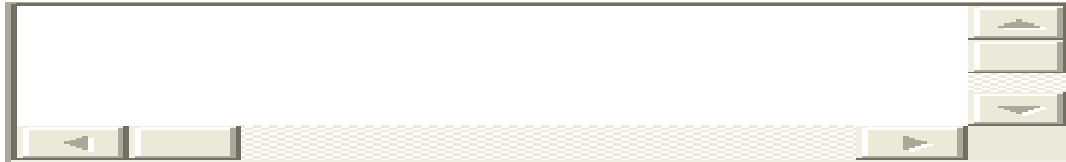
Specific Cause 4.6 **Server brings the wrong order (wrong crust, wrong toppings, etc.).**

Your Complaint Description
(Do not forget to mention the specific cause and the major cause category!)

Major Cause Category 4: **Customer Service**

Specific Cause 4.7 **Pizza is not served appropriately (plates, napkins, knives, etc.).**

Your Complaint Description
(Do not forget to mention the specific cause and the major cause category!)



Q1. Is English your first language?

- Yes (I am fluent.)
- No (But I am fluent.)
- No (But I am very good.)
- No (I am still learning.)

Q2. How experienced do you consider yourself in understanding quality-related problem causes in a production or in a service setting?

- Expert
- Very experienced
- Somewhat experienced
- Somewhat inexperienced
- Very inexperienced

Q3. How competent do you consider yourself in conceptualizing problem causes?

- Very competent (If I read a symptom description, I can see the underlying cause.)
- Relatively competent (If I read a symptom description, I may be able to figure out the underlying cause.)
- Relatively challenged (If I read a symptom description, I may not be able to figure out the underlying cause.)
- Very challenged (If I read a symptom description, I cannot see an underlying cause.)

Q4. How competent do you consider yourself in articulating (describing) problem symptoms?

- Very competent (If I read a cause, I can easily come up with a customer complaint example.)
- Relatively competent (If I read a cause, I may be able to come up with a customer complaint example.)
- Relatively challenged (If I read a cause, I may not be able to come up with a customer complaint example.)
- Very challenged (If I read a cause, I cannot come up with a customer complaint example.)

Q5. As you were providing the requested customer complaint examples, to what extent were you able to relate to a complaining customer?

- A. Very much (I was having flashbacks of dirty restaurants, badly-made pizzas, and rude servers.)
- B. To some extent (I have been through some similar situations before, so I could somewhat relate.)
- C. Very little (I could relate to some situations, but many of them made little sense to me.)
- D. Not at all (I could not imagine why a customer would ever complain in that way.)

Q6. What is your gender?

- Male
- Female

Q7. What is your age category?

- 18-20
- 21-24
- 25-34
- 35-44
- 45 and over

Q8. What is the highest education degree that you hold?

- High School

- Bachelor's
- Master's
- Ph.D

Q9. PLEASE ENTER YOUR NAME IN THE BOX.

NOTE: Your name will be used only for purposes of sending your extra credit points to your instructor. Your name will not be associated with your responses in any other way. Subsequently, your name will be deleted from this database and your responses will be processed by the researchers anonymously.

APPENDIX C

ONLINE SURVEY FOR PART II PART OF RESEARCH

Comment #1: When I went to your restaurant last night, I asked the waiter for a recommendation. He did not seem to want to help me with my decision. He was not helpful and did not offer any tips at all.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #2: I was very disappointed that the pizza in this restaurant was not very tasty. I came into the restaurant with high expectations and was let down when I noticed the condition of the food. Overall it was a disappointing visit.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #3: Exactly how old or overcooked IS this pizza? Leaving it out causes the food condition to suffer. It is either burnt or days old, it does not taste fresh. This pizza place should work on serving better pizza

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.

- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #4: The pizza didn't look like it was freshly made. We need to cook it longer to make the food processing aspect of it better.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).

- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #5: Pizza restaurants should make sure that their pizzas are enjoyed by a great number of people. Having the pizzas taste tested, could have prevented the pizza crust from being too hard. This is a violation of the food condition.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.

- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #6: The food condition was terrible. The pizza was undercooked which made it not very tasty. The pepperoni's were still frozen and the cheese wasn't even melted. This pizza was overall just disgusting.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.

- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #7: The last time I was in the bathroom at this restaurant I was afraid to touch anything. I don't think it has ever been cleaned. Any facility that has a bathroom this dirty makes you wonder what the kitchen looks like.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #8: I think they are trying to save money by being so cheap with their ingredients. The pizza barely has any cheese on it all, and the pepperoni slices are tiny.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #9: Value today is so important, and recently I must say, I felt ripped off at your Pizza restaurant. My family ordered an extra large pie for \$22.15 which we thought was pricey, but it said extra large so I let it pass. When the pizza arrived we were so shocked we took it back to the counter thinking it was a medium. We were told it was in fact the extra large size. We split the pizza three

ways and we all left hungry and 22 bucks poorer. The taste was ok, but the value was very disappointing.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #10: I took a bite of my pizza and started chewing to find that there was something wrong with the pizza dough. The dough was not mixed well and there were lumps in the dough. It made the pizza difficult to eat and not very appetizing. The way the dough was prepared was incorrect.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.

- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)

Comment #11: The pizza that i had, looks like it has been left out for a long time. When picking up a piece of pizza i could tell that it didnt break away from the cheese it just sat there like the pizza was could. After getting it off the other slice, it was so soggy i couldnt even enjou it.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.

- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #12: It made me very angry that the server ignored me and was not willing to accommodate my seating preference.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.

- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #13: Can you believe it? He brings the pizza but leaves us here without plates or silverware. The pizza's getting cold.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.

- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment # 14: The food was terrible. We ordered a cheese pizza and it came out looking more like a "sauce" pizza with a smattering of cheese scattered on the top. I was almost like the preparer ran out of cheese and was using the last of what they had.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.

- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment # 15: The crust was so lumpy. It was noticeable that the dough was not mixed well. I think that mixing the dough for a little longer will result in a better pizza. The processing just needs more time.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #16: My dining experience at the restaurant was below the quality standards, I have come to enjoy and appreciate because they prepared used a large amount of spice in my pizza. I understand you want to make this pizza taste the best it can, but please do something about this over-usage because it's turning away customers!

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #17: To provide great customer service at a restaurant a server must pay attention to a customer's order. Our business has gotten numerous complaints that servers bring the wrong order (wrong crust, wrong topping, etc.)

- **Food Condition:** Pizza is not very tasty.

- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #18: I was askin what she liked most out of the menu, she said she doesnt eat here becuae its to fatty. When i asked if anyone has liked this she said no, i already told you i dont eat hear.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.

- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #19: I've been going to this restaurant for several years and they have not ever updated the dining room. I thought it was kinda boring the first time I came but man now it's just kind of embarrassing how cheap and old this place looks.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).

- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #20: How long ago did they make this? The pizza looks old, and the texture is cold. The kitchen staff did a poor job of keeping a timely method on the process of cooking this pizza. Shame on them.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.

- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #21: I walked into the bathroom to wash my hands and the toilets are over flowing and soap is on the mirror and paper towels are everywhere. your bathrooms are not clean and need attention from the janitor.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions

- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #22: The food i ordered isn't right. I didn't order all these topping on my salad and im allergic to most of them. They need to write down my exact order so they dont send someone to the hospital! This is very seriouse and makes me upset!

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

comment #23: I cannot believe how dirty the tables and floor were. The workers have not picked up after any of the customers. They need to focus on their customers.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #24: The employee's uniforms are dirty the facility should offer some kind of laundry service

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.

- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #25: I was not pleased with the service i recieved. I was ignored several times when requesting something from employees. When they finally did respond they did so unprofessionally.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.

- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #26: The food processing methods were poor. I saw one of the employees cleaning the floors with bleach and some of that bleach bounced on to the pizza. Unacceptable.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).

- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #27: The food processing was terrible. The pizza was way to overcooked to serve to customers.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.

- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Customer #28: When the waiter finally came around, i asked her if i could get a refill on my drink, she gave me a go to hell look, then proceeded to take 10 minutes and pretty much threw my drink at me. Needless to say i didnt drink it.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.

- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #29: The pizza toppings are not cut up but served on top of the pizza in much larger pieces. One slice of onion does not cover a pizza well. Lots of little piece cover the pizza so you have onion in all your bites, they are stingy with their ingredients.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #30: The pizza has too much spices. It is too spicy to eat, I feel like I am going to get sick.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #31: When watching the tv and saw an amazing pizza on tv, i decided to order one, when i got my pizza it wasnt even round and there was flour not mixed properly into the dough. It looked like a huge glob of dough and ingrediants. They should of left the dough a little longer in the mixer.

- **Food Condition:** Pizza is not very tasty.

- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #32: I caught a glimpse of the kitchen, and I am disturbed! It looks dirty, messy, and disorganized. I fear that by the condition of the kitchen, what is the food standards for this facility?

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.

- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #33: I do not want my food cooked in that kitchen. The Pizza restaurant's facility is terrible. The kitchen area looks disturbingly dirty, disorganized, and messy.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under- or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.

- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #34: The employees were polite as we first walked in. However, it didn't take long for them to change to being rude and hustling customers to hurry up. To say the least, I was at my limit! There was probably smoke coming out of my ears I was so mad!

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.

- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #35: The pizza I got tasted absolutely terrible! It tasted like Windex! I believe there was some cleaner used by the food. If I wanted the Windex taste, I would have gone to Dominos.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions

- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #36: I waited 45 minutes for my food. I ordered a slice of cheese pizza. The place was empty.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #37: the bathrooms are really nasty and i could not see myself using them. i told the manager and he seemed no to care at all.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #38: Wings are cold. Bad food processing method

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #39: When I walked to the table to be seated I wanted to get wet wipes out and clean it. everything looked so dirty, and i will never come back.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.

- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #40: I can literally smell the stains from your shirt across the counter. It's really disgusting. The restaurant owner arranges for facility clean up, but you need to wash your own uniform...pronto.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.

- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #41: The customer service is not paying attention to the dining needs of the customer. There needs to be plates, napkins and utensils before the food arrives. The server does not care enough to accomodate to the customers dining experience.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.

- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #42: the server had poor customer service with me. i sat there for far too long waiting for assistance and then i when i finally did i had to remind them 3 times what i needed because he kept forgetting.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.

- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #43: The pizza served tasted bad. It didn't have enough spices on it. The food quality was subpar to what was expected. Many people left their pizza on their plate, and left in a huff from the restaurant.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)

- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #44: When I used to come into your restaurants I was greeted by clean tables and bathrooms. Recently, however, I have found that the facilities are somewhat less clean to the point that I would actually warn people away from dining at your restaurant. Please tell your cleaning staff that they need to spend a bit more time cleaning the tables, as the tables looked unclean and still had trash from previous customers sitting on them

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under- or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #45: I ordered a milkshake over 25 minutes ago. Where is my milkshake!? It should not take this long to get a stinking milkshake.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #46: The server brought out a completely different pizza than what we ordered! He must have not been paying attention to what we were saying, and he didnt even apologize when it happened. He had absolutely no customer service skills.

- **Food Condition:** Pizza is not very tasty.

- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #47: One of the problems with your facility is that your employees ALL have dirty uniforms. And it seemed like a couple of them hadn't been washed in a while. Either get more responsible people who know how to clean their clothing or clean their uniforms for them, because some of them were very stinky!

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).

- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #48: Great food, wonderful service, but the decor is a bit spartan. Even a few basics like glass cups, table clothes and maybe some paint on the walls would be a welcome addition. No one like to feel lik they are eating in a mess hall. Just dissapointed the environment did not match upto the food quality.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).

- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #49: The wait time in the pizza shop was horrendous. I waited nearly two hours. What could possibly take them so long?

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.

- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #50: Some of our Host have been seating our customes in the smoking area when they specifically stated they want to be in the non smoking area. Our customer service must get better we do not want to lose business.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.

- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #51: Everyone in this parlor is upset because The Pizza in this parlor is rediculously small. the large is the size of a personal pizza. theres no way i can feed my family of four when you serve food in this condition.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)

- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #52: The pizza kitchen area was disgusting. You can tell there is no organization back there, ingredients and cooking supplies were strewn all over the kitchen.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #53: I can't believe how undercooked my pizza was. The dough was still chewy, and my cheese was not melted. Please learn how to cook your pizza's a little longer.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #54: The pizza shop has good pizza but not a good atmosphere. It is very plain and undecorated. I would like a few decorations and maybe some music while eating.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).

- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #55: I asked the host for some extra chairs in order to seat my large family. The host refused to accommodate my reasonable request. My grandparents had to sit at separate table even though there was room to move in extra chairs. That's not good customer service.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.

- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #56: When I eat at the pizza restaurant, I never receive the desired amount of utensils and plates needed. When I ask the server for these things, it takes forever to get them, and it's very upsetting. I believe the restaurant could better its service and reputation if the server gives the customers the right amount of utensils they need, even if it's more than enough.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under- or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).

- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #57: I enjoy eating this pizza, but it's gone all too quickly... I'm waiting for the main course. Fun-sized pizza isn't very fun.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.

- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #58: I feel as if most of the people who are taking my order all are on their first day. When I ask for recommendations, they don't seem no know what's good. Makes me a little nervous about buying food at this location is the workers don't even eat it.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.

- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #59: I waited 15 minutes just to get a refill. Servers should constantly check on who they are waiting on. Though it is mainly a lost of money for you it could possibly be a lost of business for us.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or tooppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #60: I was upset with the methods you used to prepare your food. This pizza was very undercooked and was not satisfying at all. I would recommend different food processing methods.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #61: The employees tonight at the Pizza Restaurant were extremely rude and impolite! I can't believe the nerve of the workers here. This company's customer service needs to be revamped.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.

- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #62: There is a hair in my pizza! I do not feel like this pizza is safe to eat, due to this contamination. I feel like the food processing methods here need to be improved to prevent such contamination in the future.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.

- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).
- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Comment #63: I waited forever to get my pizza at this place and they hand me a pizza with crust that was as hard as a rock. It tells me that they forgot about the pizza and the condition of it was not pleasing.

- **Food Condition:** Pizza is not very tasty.
- **Food Condition:** Pizza size is too small.
- **Food Condition:** Quantity of pizza ingredients such as cheese or toppings is too small (restaurant is being too stingy).
- **Food Condition:** Pizza has either too much or too little amount of spices.
- **Food Condition:** Pizza crust is too hard.
- **Food Processing Methods:** Pizza is under-or overcooked.
- **Food Processing Methods:** Pizza was not safe to eat (possible contamination).
- **Food Processing Methods:** Pizza does not look freshly made (cold pizza).

- **Food Processing Methods:** Pizza dough was not mixed well (has lumps inside).
- **Facility:** The restaurant's entire eating area looks dirty.
- **Facility:** Employee uniforms are dirty.
- **Facility:** Kitchen area looks dirty, disorganized, messy.
- **Facility:** Environment is too plain, too undecorated.
- **Facility:** Bathrooms are not clean.
- **Customer Service:** Employees are impolite or rude.
- **Customer Service:** When I need help, it takes too long for the server to notice.
- **Customer Service:** Server is not helpful while I try to make my order decisions
- **Customer Service:** Host is not willing to accommodate my seating preference.
- **Customer Service:** My order takes too long to be served.
- **Customer Service:** Server brings the wrong order (wrong crust, wrong toppings, etc.)
- **Customer Service:** Pizza is not served appropriately (plates, napkins, knives, etc.).

Q1. Is English your first language?

- Yes (I am fluent.)
- No (But I am fluent.)
- No (But I am very good.)
- No (I am still learning.)

Q2. How experienced do you consider yourself in understanding quality-related problem causes in a production or in a service setting?

- Expert
- Very experienced
- Somewhat experienced
- Somewhat inexperienced
- Very inexperienced

Q3. How competent do you consider yourself in conceptualizing problem causes?

- Very competent (If I read a symptom description, I can see the underlying cause.)
- Relatively competent (If I read a symptom description, I may be able to figure out the underlying cause.)
- Relatively challenged (If I read a symptom description, I may not be able to figure out the underlying cause.)
- Very challenged (If I read a symptom description, I cannot see an underlying cause.)

Q4. How competent do you consider yourself in articulating (describing) problem symptoms?

- Very competent (If I read a cause, I can easily come up with a customer complaint example.)
- Relatively competent (If I read a cause, I may be able to come up with a customer complaint example.)
- Relatively challenged (If I read a cause, I may not be able to come up with a customer complaint example.)
- Very challenged (If I read a cause, I cannot come up with a customer complaint example.)

Q5. As you were reading the customer comments, to what extent were you able to relate to a complaining customer?

- A. Very much (I was having flashbacks of dirty restaurants, badly-made pizzas, and rude servers.)
- B. To some extent (I have been through some similar situations before, so I could somewhat relate.)
- C. Very little (I could relate to some situations, but many of them made little sense to me.)
- D. Not at all (I could not imagine why a customer would ever complain in that way.)

Q6. What is your gender?

- Male
- Female

Q7. What is your age category?

- 18-20

- 21-24
- 25-34
- 35-44
- 45 and over

Q8. What is the highest education degree that you hold?

- High School
- Bachelor's
- Master's
- Ph.D

Q9. PLEASE ENTER YOUR NAME IN THE BOX.

NOTE: Your name will be used only for purposes of sending your extra credit points to your instructor. Your name will not be associated with your responses in any other way. Subsequently, your name will be deleted from this database and your responses will be processed by the researchers anonymously.

APPENDIX D

IRB10480 INFORMATION NOTICE

University of North Texas Institutional Review Board Information Notice

Before agreeing to participate in this research study, it is important that you read and understand the following explanation of the purpose, benefits and risks of the study and how it will be conducted.

Title of Study: A Conceptualization-Articulation Study for Quality Problems – Part 1 _____
Principal Investigator: _Dr. Nicholas Evangelopoulos, University of North Texas (UNT) Department of ITDS.

Purpose of the Study: This study is designed to gather anticipated customer complaint statements related to quality problems in a fictitious restaurant setting. For each provided quality problem category, the subjects are asked to use their own words and provide examples of how they expect customers to word their complaints. The purpose of this study is to generate customer complaint data, in free text form. The textual data will be analyzed for automatic classification and analysis results will be compared against human categorization.

Study Procedures: You will be asked to fill out an online survey. The survey is expected to take approximately 30 minutes. An alternative non-research activity with equivalent time and effort is also offered. The alternative assignment asks you to submit a 1½-page report providing one example for each quality problem cause category listed on a cause-and-effect diagram that appears on the course textbook (Metro Delivery Service case study, pp. 341-342 of the DSCI 2710 Business Statistics textbook, 2009 courseware edition). The alternative assignment is also worth the same points of extra credit.

Foreseeable Risks: No foreseeable risks are involved in this study.

Benefits to the Subjects or Others: This study is asking the participants to articulate quality problem causes. The cause categories are organized in a cause-and-effect diagram. As such, the study is directly related to course material in Decision Sciences. With respect to instructional benefits, the study is expected to give participants a chance to better understand how service quality problems appear in a restaurant setting. With respect to research-related benefits, the results are expected to benefit the principal investigator's field of study. The text mining techniques that will be applied on data generated by this study could also apply to the analysis of real customer comments. For example, many restaurants have a suggestion box soliciting customer feedback. This study will help us determine effective ways to analyze this type of comment data.

Compensation for Participants: The course coordinator and your instructor have agreed to award DSCI 2710 student participants with 10 extra credit points on the 800-point course grading

scale and would like to have the names of the students who would have completed the survey. The alternative non-research assignment is also worth 10 points of extra credit. For DSCI 4520/5240 participants, there will be a 1% extra credit bonus for participation.

Procedures for Maintaining Confidentiality of Research Records: All paperwork resulting from this research will only be handled by the people assigned to work on this research, which are: Dr. Nicholas Evangelopoulos, and Ms. Leticia Anaya. Your course coordinator and/or your instructor have agreed to award participants with extra credit points and would like to have the names of the students who would have completed the survey. The survey has been adjusted to accommodate the gathering of these names. However, as soon as the instructors get the lists of students who completed the survey, the survey responses will be coded and the names will be deleted to preserve anonymity.

Subsequently, survey responses and names of participants will be kept in separate places. All instruments will be maintained under lock and key for a period of three years. The survey responses will be analyzed using published text mining and statistical techniques. No names will be posted on coded surveys, which will be administered in a way that maintains confidentiality of individual information at all times. Outcomes of this study that may be published or presented will only include aggregate results and statistical summaries.

Questions about the Study: If you have any questions about the study, you may contact *Dr. Nicholas Evangelopoulos* at telephone number (940) 565-3056 or via e-mail at Nick.Evangelopoulos@unt.edu.

Review for the Protection of Participants: This research study has been reviewed and approved by the UNT Institutional Review Board (IRB). The UNT IRB can be contacted at (940) 565-3940 with any questions regarding the rights of research subjects.

Research Participants' Rights:

Your participation in the online survey indicates that you have read all of the above and that you confirm all of the following:

- You understand the possible benefits and the potential risks and/or discomforts of this study.
- You understand that you do not have to take part in this study, and your refusal to participate or your decision to withdraw will involve no penalty or loss of rights or benefits. The study personnel may choose to stop your participation at any time.
- You understand why the study is being conducted and how it will be performed.
- You understand your rights as a research participant and you voluntarily consent to participate in this study.
- Your decision to participate or to withdraw from the study will have no effect on your standing in this course or your course grade. You can receive the same extra credit by doing the alternative assignment.
- You may download and keep a copy of this form.

Approved by IRB from 11/9/10 to 11/8/11

APPENDIX E
NOVEMBER 8, 2010, IRB APPLICATION

Conceptualization-Articulation, Nov-8-2010, IRB Application

IRB 10480 – Revised after initial review

**Expedited or Full Board
Review Application**

For IRB Use Only	
File Number:	
Approval	

University of North Texas Institutional Review Board
OHRP Federalwide Assurance: FWA00007479

Save this file as a Word document on your computer, answer all questions completely within Word, and submit it along with all supplemental documents to the IRB Office as described in the Electronic Submission Checklist on page 5.

Type only in the **yellow** fields, and closely follow all stated length limits. Handwritten forms will not be accepted.

1. Title of Study
Must be identical to the title of any related internal or external grant proposal.
A CONCEPTUALIZATION-ARTICULATION STUDY FOR QUALITY PROBLEMS – PART 1

2. Investigator Information		
Must be: (a) a UNT faculty member; and (b) the same person as the Principal Investigator named in any related proposal for external or internal funding.		
Nicholas	Evangelopoulos	Nick.Evangelopoulos@unt.edu
First Name	Last Name	E-mail Address
ITDS	COBA	B A 302 G
UNT Department	UNT Building	Room Number
940 565 -3056	940)565-4935	
Office Phone Number	Fax Number	

3. Co-Investigator Information		
Must be a UNT faculty member.		
First Name	Last Name	E-mail Address
UNT Department	Title	

4. Key Personnel
List the name of all other Key Personnel who are responsible for the design, conduct, or reporting of the study (including recruitment or data collection).

Leticia H. Anaya (Lanaya@unt.edu), PhD COBA Student and Lecturer in ETEC Dept.

NIH IRB Training

Have you, any Co-Investigator, and all Key Personnel completed the required NIH IRB training course (“Protecting Human Research Participants”) and electronically submitted a copy of the completion certificate to untirb@unt.edu?

Yes No

If “No,” this training is required for all Key Personnel before your study can be approved. This free on-line course may be accessed at: <http://phrp.nihtraining.com>

5. Funding Information (If applicable)

Provide the proposal number or project ID number for any external funding or the account number for any internal funding for this project

Not Applicable

6. Purpose of Study

In no more than half a page, briefly state the purpose of your study in **lay language**, including the research question(s) you intend to answer. A brief summary of what you write here should be included in the Informed Consent document.

This study is designed to gather anticipated customer complaint statements related to quality problems in a fictitious restaurant setting. For each provided quality problem category, the subjects are asked to use their own words and provide examples of how they expect customers to word their complaints. The purpose of this study is to generate customer complaint data, in free text form. The textual data will be analyzed for automatic classification and analysis results will be compared against human categorization.

7. Previous Research

In no more than half a page, summarize previous research leading to the formulation of this study, including any past or current research conducted by the Investigator or key personnel that leads directly to the formulation of this study (including citations and references.)

The PI has conducted and published related research in the past, including research that analyzed customer comments with the use of Latent Semantic Analysis (Evangelopoulos 2007), and published abstracts of research articles (Sidorova et al. 2008). To address various methodological issues that revolve around the employment of Latent Semantic Analysis, the PI has a paper recently accepted for

publication in a special journal issue on quantitative methods (Evangelopoulos et al. 2011). The proposed study is the PI's first attempt in employing Latent Dirichlet Allocation. The study is closely related to Leticia Anaya's doctoral dissertation, supervised by the PI (Ms. Anaya is listed in this application under Key Personnel). We consider this study to be interesting and have high publication potential.

References

Evangelopoulos, N. (2007), "Analyzing Free-Text Customer Feedback," Proceedings of the 38th Annual Meeting of the Decision Sciences Institute, Phoenix, AZ, November 17-20, 2007, pp. 3731-3736.
 Evangelopoulos, N., Zhang, X, and Prybutok, V. (2011), "Latent Semantic Analysis: Five Methodological Considerations." European Journal of Information Systems, special issue on Quantitative Research Methods. Forthcoming.
 Sidorova, A., Evangelopoulos, N., Valacich, J., and Ramakrishnan, T. (2008), "Uncovering the Intellectual Core of the Information Systems Discipline," MIS Quarterly, 32(3), pp. 467-482.

8. Recruitment of Participants
Describe the projected number of subjects.
700-800 students
Describe the population from which subjects will be recruited (including gender, racial/ethnic composition, and age range).
DSCI2710, DSCI4520, and DSCI5240 students. These are undergraduate & graduate UNT students enrolled at traditional or online sections. Typical UNT student gender/age/racial/ethnic composition is expected.
Describe how you will recruit the subjects.
Students are enrolled in relevant classes (DSCI 2710, 4520, and 5240). Survey participation will be announced by the course instructors as an extra credit opportunity.
Have you attached a copy of all recruitment materials?
<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

9. Vulnerable Populations
Please identify any vulnerable populations who will be targeted for participation in this study:
<input type="checkbox"/> Children (under 18 years of age) <input type="checkbox"/> Pregnant Women <input type="checkbox"/> Prisoners <input type="checkbox"/> Mentally Impaired or Mentally Retarded
If any boxes are checked, describe any special precautions to be taken in your study due to the inclusion of these populations.
We do not expect any participant to belong to these special populations.

10. Location of Study
Identify all locations where the study will be conducted.
UNT College of Business (Denton, Texas). DSCI 2710, 4520, and 5240 classrooms.
For data collection sites other than UNT, have you obtained and documented permission to recruit subjects and collect data at each site?
<input type="checkbox"/> Yes <input type="checkbox"/> No (N/A)

11. Informed Consent

Describe the steps for obtaining the subjects' informed consent (by whom, where, when, etc.).

The survey will be taken by the participants as an internet survey. The informed consent page will therefore be posted on the Web as an Informed Consent Notice.

12. Informed Consent Forms

Written Informed Consent Forms to be signed by the subject are required for most research projects with human participants (exceptions include telephone surveys, internet surveys, and other circumstances where the subject is not present; an Informed Consent Notice may be substituted). Templates for creating consent forms are located on the IRB website at <http://research.unt.edu/ors/compliance/human.htm>. **All informed consent documents you plan to use must be submitted before IRB review can begin.**

13. Foreign Languages

Will your study involve the use of any language other than English for Informed Consent forms, data collection instruments, or recruitment materials?

Yes No

If "Yes," after the IRB has notified you of the approval of the English version of your forms, you must then submit the foreign language versions along with a back-translation for each. Specify all foreign languages below:

14. Data Collection

Which methods will you use to collect data?

- | | |
|---|--|
| <input type="checkbox"/> Interviews | <input checked="" type="checkbox"/> Internet Surveys |
| <input type="checkbox"/> Surveys | <input type="checkbox"/> Review of Existing Records |
| <input type="checkbox"/> Focus Groups | <input type="checkbox"/> Observation |
| <input type="checkbox"/> Assessment Instruments | <input type="checkbox"/> Other – Please list below. |

An online survey has been created and the URL link to that survey is:

https://untbusiness.qualtrics.com/SE/?SID=SV_7TH2q0glsSUwVUg&Preview=Survey&BrandID=untbusiness

Students will be provided the link to the survey. The instructors have agreed to give extra credit for the completion of this survey.

Have you attached a copy of all data collection instruments, interview scripts, and intervention protocols to be used?

Yes No

Provide a list of all data collection instruments below. Attach a copy and label each instrument.

Conceptualization-Articulation1_SurveyInstrument (see attached)

What is the estimated time for a subject's participation in each study activity (including time per session and total number of sessions)?

It is expected that the survey will take approximately 30 minutes to complete. There will only be one session.

15. HIPAA

Will your study involve obtaining individually identifiable health information from health care plans, health care clearinghouses, or health care providers?

Yes No Not Applicable

If "Yes," describe the procedures you will use to comply with the HIPAA Privacy Rule. (For more information about HIPAA, see the HIPAA Guidance page on the IRB website at <http://research.unt.edu/ors/compliance/hipaa.htm>.)

16. Compensation

Describe any compensation subjects will receive for participating in the study. Include the timing for payment and any conditions for receipt of such compensation. If extra credit for a course is offered, an alternative non-research activity with equivalent time and effort must also be offered.

The DSCI 2710 course coordinator and the respective section instructors have agreed to give extra credit to their DSCI 2710 students. The extra credit will be 10 points on the 800-point course grading scale. An alternative non-research activity with equivalent time and effort is also offered. The alternative assignment asks the students to submit a 1½-page report providing one example for each quality problem cause category listed on a cause-and-effect diagram that appears on the course textbook. (Metro Delivery Service case study, pp. 341-342 of the DSCI 2710 Business Statistics textbook, 2009 courseware edition). The alternative assignment is also worth 10 points of extra credit. Similar arrangements and the same alternative non-research assignment hold for DSCI 4520/5240, where a 1% extra credit bonus will be awarded.

17. Risks and Benefits

Describe any foreseeable risks to subjects presented by the proposed study and the precautions you will take to minimize such risks.

There are no foreseeable risks.

Describe the anticipated benefits to subjects or others (including your field of study).

This study is asking the participants to articulate quality problem causes. The cause categories are organized in a cause-and-effect diagram. As such, the study is directly related to course material in Decision Sciences (for example, DSCI 2710 is a business statistics course that includes a quality control chapter). With respect to instructional benefits, the study is expected to giving participants a chance to better understand how service quality problems appear in a restaurant setting. With respect to research-

related benefits, the results are expected to benefit the PI's and the participating doctoral student's field of study. The application of text mining techniques similar to those involved in the proposed study can be extended to apply to automatic categorization problems in a variety of settings.

18. Confidentiality

Describe the procedures you will use to maintain the confidentiality of any personally identifiable data.

All paperwork resulting from this research will only be handled by the people assigned to work on this research, which are: Dr. Nicholas Evangelopoulos, and Ms. Leticia Anaya. The DSCI 2710 course coordinator and the DSCI 2710, 4520, and 5240 course instructors have agreed to award participants with extra credit points and would like to have the names of the students who would have completed the survey. The survey has been adjusted to accommodate the gathering of these names. However, as soon as the instructors get the lists of students who completed the survey, the survey responses will be coded and the names will be deleted to preserve anonymity. Subsequently, survey responses and names of participants will be kept in separate places. All instruments will be maintained under lock and key for a period of three years. The survey responses will be analyzed using published text mining and statistical techniques. No names will be posted on coded surveys, which will be administered in a way that maintains confidentiality of individual information at all times. Outcomes of this study that may be published or presented will only include aggregate results and statistical summaries.

Please specify where your research records will be maintained, any coding or other steps you will take to separate participants' names/identities from research data, and how long you will retain personally identifiable data in your research records. Federal IRB regulations require that the investigator's research records be maintained for 3 years following the end of the study.

The names will be separated from the corresponding survey answers at an early stage. Subsequently, Ms. Leticia Anaya will code survey responses with the use of random re-ordering and system-generated response ID numbers, so that anonymity of research data is maintained. Research records will be maintained under lock and key for a period of three years in the researchers' offices.

19. Publication of Results

Please identify all methods in which you may publicly disseminate the results of your study.

- | | |
|--|---|
| <input checked="" type="checkbox"/> Academic journal | <input checked="" type="checkbox"/> A thesis or dissertation for one of your students |
| <input checked="" type="checkbox"/> Academic conference paper or public poster session | <input type="checkbox"/> Other - Please list below.
(E.g. website or blog) |
| <input type="checkbox"/> Book or chapter | |

Investigator Signature

I certify that the information in this application is complete and accurate. I agree to conduct this study in accordance with the UNT IRB Guidelines and the study procedures and forms approved by the UNT IRB. I understand that I cannot initiate any contact with potential human subjects until I have received written UNT IRB approval.

11/8/2010



Nicholas
Evangelopoulos

Signature of Investigator

Date

Signature of Co- Investigator

Date

Electronic Submission Checklist

1. Print and sign this page and then scan the signed document.
2. Attach all supplementary documents, including:
 - a. Copies of all NIH Training completion certificates not previously submitted to the IRB Office;
 - b. A copy of any proposal for internal or external funding for this study;
 - c. A copy of all recruitment materials;
 - d. A copy of all informed consent forms; and
 - e. A copy of all data collection instruments, interview scripts and intervention protocols.
3. E-mail the application (including this Signature Page) and all supplementary documents to untirb@unt.edu. Please insert "Expedited or Full Board Review" in the subject line of your email.

Contact Shelia Bourns at Shelia.Bourns@unt.edu for any questions about completion of your application.

APPENDIX F

IRB 10504 INFORMATION NOTICE

University of North Texas Institutional Review Board Information Notice

Before agreeing to participate in this research study, it is important that you read and understand the following explanation of the purpose, benefits and risks of the study and how it will be conducted.

Title of Study: A Conceptualization-Articulation Study for Quality Problems – Part 2

Principal Investigator: Dr. Nicholas Evangelopoulos, University of North Texas (UNT) Department of ITDS.

Purpose of the Study: In this study, participants are asked to act as customer relationship managers and evaluate customer complaint statements related to quality problems in a fictitious restaurant setting. The participants are presented with 63 customer comments and asked to categorize each one of them under one of 21 listed categories. The purpose of this study is to assess human performance in categorizing data in free text form. The same textual data will be analyzed using automatic classification algorithms and analysis results will be compared against human categorization.

Study Procedures: You will be asked to fill out an online survey. The survey is expected to take approximately 30 minutes. An alternative non-research activity with equivalent time and effort is also offered. The alternative assignment asks you to submit a 1½-page report providing one example for each quality problem cause category listed on a cause-and-effect diagram that appears on the DSCI 2710 textbook (Metro Delivery Service case study, pp. 341-342 of the DSCI 2710 Business Statistics textbook, 2009 courseware edition). A handout with details on the alternative assignment will be posted on your course Web site. The alternative assignment is also worth the same points of extra credit.

Foreseeable Risks: No foreseeable risks are involved in this study.

Benefits to the Subjects or Others: This study is asking the participants to generalize specific customer comments and categorize them under broad quality problem causes. The cause categories are organized in a cause-and-effect diagram. As such, the study is directly related to course material in Decision Sciences. With respect to instructional benefits, the study is expected to give participants a chance to better understand how service quality problems appear in a restaurant setting. With respect to research-related benefits, the results are expected to benefit the PI's and the participating doctoral student's field of study. The text mining techniques used in this study could also apply to the analysis of real customer comments. For example, many restaurants have a suggestion box soliciting customer feedback. This study will help us determine effective ways to analyze this type of comment data.

Compensation for Participants: The DSCI 4520, 5240, and 5320 instructors have agreed to give extra credit to their students. An alternative non-research activity with equivalent content, time and effort is also offered. The alternative non-research assignment is also worth the same extra credit.

Procedures for Maintaining Confidentiality of Research Records: All paperwork resulting from this research will only be handled by the people assigned to work on this research, which are: Dr. Nicholas Evangelopoulos, and Ms. Leticia Anaya. Your instructor has agreed to award participants with extra credit points and would like to have the names of the students who would have completed the survey. The survey has been adjusted to accommodate the gathering of these names. However, as soon as the instructors get the lists of students who completed the survey, the survey responses will be coded and the names will be deleted to preserve anonymity. Subsequently, survey responses and names of participants will be kept in separate places. All instruments will be maintained under lock and key for a period of three years. The survey responses will be analyzed using published text mining and statistical techniques. No names will be posted on coded surveys, which will be administered in a way that maintains confidentiality of individual information at all times. Outcomes of this study that may be published or presented will only include aggregate results and statistical summaries.

Questions about the Study: If you have any questions about the study, you may contact *Dr. Nicholas Evangelopoulos* at telephone number (940) 565-3056 or via e-mail at Nick.Evangelopoulos@unt.edu.

Review for the Protection of Participants: This research study has been reviewed and approved by the UNT Institutional Review Board (IRB). The UNT IRB can be contacted at (940) 565-3940 with any questions regarding the rights of research subjects.

Research Participants' Rights:

Your participation in the online survey indicates that you have read all of the above and that you confirm all of the following:

- You understand the possible benefits and the potential risks and/or discomforts of this study.
- You understand that you do not have to take part in this study, and your refusal to participate or your decision to withdraw will involve no penalty or loss of rights or benefits. The study personnel may choose to stop your participation at any time.
- You understand why the study is being conducted and how it will be performed.
- You understand your rights as a research participant and you voluntarily consent to participate in this study.
- Your decision to participate or to withdraw from the study will have no effect on your standing in this course or your course grade. You can receive the same extra credit by doing the alternative assignment.
- You may download and keep a copy of this form.

IRB APPROVAL 12/7/10 to 12/6/11

APPENDIX G
NOVEMBER 30, 2010 IRB APPLICATION

Conceptualization-Articulation Study Part II

Expedited or Full Board Review Application

University of North Texas Institutional Review Board
OHRP Federalwide Assurance: FWA00007479

For IRB Use Only	
File Number:	
Approval	

Save this file as a Word document on your computer, answer all questions completely within Word, and submit it along with all supplemental documents to the IRB Office as described in the Electronic Submission Checklist on page 5.

Type only in the **yellow** fields, and closely follow all stated length limits. Handwritten forms will not be accepted.

1. Title of Study
Must be identical to the title of any related internal or external grant proposal.
A CONCEPTUALIZATION-ARTICULATION STUDY FOR QUALITY PROBLEMS – PART 2

2. Investigator Information		
Must be: (a) a UNT faculty member; and (b) the same person as the Principal Investigator named in any related proposal for external or internal funding.		
Nicholas	Evangelopoulos	Nick.Evangelopoulos@unt.edu
First Name	Last Name	E-mail Address
ITDS	COBA	B A 302 G
UNT Department	UNT Building	Room Number
940 565 -3056	940)565-4935	
Office Phone Number	Fax Number	

3. Co-Investigator Information		
Must be a UNT faculty member.		
First Name	Last Name	E-mail Address
UNT Department	Title	

4. Key Personnel
List the name of all other Key Personnel who are responsible for the design, conduct, or reporting of the study (including recruitment or data collection).

Leticia H. Anaya (Lanaya@unt.edu), PhD COBA Student and Lecturer in ETEC Dept.

NIH IRB Training

Have you, any Co-Investigator, and all Key Personnel completed the required NIH IRB training course (“Protecting Human Research Participants”) and electronically submitted a copy of the completion certificate to untirb@unt.edu?

Yes No

If “No,” this training is required for all Key Personnel before your study can be approved. This free on-line course may be accessed at: <http://phrp.nihtraining.com>

5. Funding Information (If applicable)

Provide the proposal number or project ID number for any external funding or the account number for any internal funding for this project

Not Applicable

6. Purpose of Study

In no more than half a page, briefly state the purpose of your study in **lay language**, including the research question(s) you intend to answer. A brief summary of what you write here should be included in the Informed Consent document.

In this study, participants are asked to act as customer relationship managers and evaluate customer complaint statements related to quality problems in a fictitious restaurant setting. The participants are presented with 63 customer comments and asked to categorize each one of them under one of 21 listed categories. The purpose of this study is to assess human performance in categorizing data in free text form. The same textual data will be analyzed using automatic classification algorithms and analysis results will be compared against human categorization.

7. Previous Research

In no more than half a page, summarize previous research leading to the formulation of this study, including any past or current research conducted by the Investigator or key personnel that leads directly to the formulation of this study (including citations and references.)

The PI has conducted and published related research in the past, including research that analyzed customer comments with the use of Latent Semantic Analysis (Evangelopoulos 2007), and published abstracts of research articles (Sidorova et al. 2008). To address various methodological issues that revolve around the employment of Latent Semantic Analysis, the PI has a paper recently accepted for

publication in a special journal issue on quantitative methods (Evangelopoulos et al. 2011). The proposed study is the PI's first attempt in employing Latent Dirichlet Allocation. The study is closely related to Leticia Anaya's doctoral dissertation, supervised by the PI (Ms. Anaya is listed in this application under Key Personnel). We consider this study to be interesting and have high publication potential.

References

Evangelopoulos, N. (2007), "Analyzing Free-Text Customer Feedback," Proceedings of the 38th Annual Meeting of the Decision Sciences Institute, Phoenix, AZ, November 17-20, 2007, pp. 3731-3736.
 Evangelopoulos, N., Zhang, X, and Prybutok, V. (2011), "Latent Semantic Analysis: Five Methodological Considerations." European Journal of Information Systems, special issue on Quantitative Research Methods. Forthcoming.
 Sidorova, A., Evangelopoulos, N., Valacich, J., and Ramakrishnan, T. (2008), "Uncovering the Intellectual Core of the Information Systems Discipline," MIS Quarterly, 32(3), pp. 467-482.

8. Recruitment of Participants
Describe the projected number of subjects.
60-70 students
Describe the population from which subjects will be recruited (including gender, racial/ethnic composition, and age range).
DSCI 4520, DSCI 5240, and DSCI 5320 students. These are undergraduate & graduate UNT students enrolled at traditional or online sections. Typical UNT student gender/age/racial/ethnic composition is expected.
Describe how you will recruit the subjects.
Students are enrolled in relevant classes (DSCI 4520, 5240, and 5320). Survey participation will be announced by the course instructors as an extra credit opportunity.
Have you attached a copy of all recruitment materials?
<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

9. Vulnerable Populations
Please identify any vulnerable populations who will be targeted for participation in this study:
<input type="checkbox"/> Children (under 18 years of age) <input type="checkbox"/> Pregnant Women <input type="checkbox"/> Prisoners <input type="checkbox"/> Mentally Impaired or Mentally Retarded
If any boxes are checked, describe any special precautions to be taken in your study due to the inclusion of these populations.
We do not expect any participant to belong to these special populations.

10. Location of Study
Identify all locations where the study will be conducted.
UNT College of Business (Denton, Texas). DSCI 4520, 5240, and 5320 classrooms.
For data collection sites other than UNT, have you obtained and documented permission to recruit subjects and collect data at each site?
<input type="checkbox"/> Yes <input type="checkbox"/> No (N/A)

Conceptualization-Articulation2_SurveyInstrumentTemplate (see attached)
Conceptualization-Articulation2_SurveyInstrument_List-of-comments (see attached)

Please note: Each survey has 63 comments. There are 16 versions of the survey, totaling 1,008 comments. The attached Survey Instrument Template document shows only one sample comment. The attached List-of-comments document provides the full list of 1,008 comments. The 1,008 comments were submitted anonymously by participants of a previous study conducted by the same PI in November 2010 (UNT IRB case #10480)

What is the estimated time for a subject's participation in each study activity (including time per session and total number of sessions)?

It is expected that the survey will take approximately 30 minutes to complete. There will only be one session.

15. HIPAA

Will your study involve obtaining individually identifiable health information from health care plans, health care clearinghouses, or health care providers?

Yes No Not Applicable

If "Yes," describe the procedures you will use to comply with the HIPAA Privacy Rule. (For more information about HIPAA, see the HIPAA Guidance page on the IRB website at <http://research.unt.edu/ors/compliance/hipaa.htm>.)

16. Compensation

Describe any compensation subjects will receive for participating in the study. Include the timing for payment and any conditions for receipt of such compensation. If extra credit for a course is offered, an alternative non-research activity with equivalent time and effort must also be offered.

The DSCI 4520, 5240, and 5320 instructors have agreed to give a 1% extra credit to their students. An alternative non-research activity with equivalent content, time and effort is also offered. The alternative assignment asks the students to submit a 1½-page report providing one example for each quality problem cause category listed on a cause-and-effect diagram that appears on the DSCI 2710 course textbook. (Metro Delivery Service case study, pp. 341-342 of the DSCI 2710 Business Statistics textbook, 2009 courseware edition). The alternative assignment is also worth 1% of extra credit.

17. Risks and Benefits

Describe any foreseeable risks to subjects presented by the proposed study and the precautions you will take to minimize such risks.

There are no foreseeable risks.

Describe the anticipated benefits to subjects or others (including your field of study).

This study is asking the participants to generalize specific customer comments and categorize them under broad quality problem causes. The cause categories are organized in a cause-and-effect diagram. As such, the study is directly related to course material in Decision Sciences. With respect to instructional benefits, the study is expected to give participants a chance to better understand how

service quality problems appear in a restaurant setting. With respect to research-related benefits, the results are expected to benefit the PI's and the participating doctoral student's field of study. The application of text mining techniques similar to those involved in the proposed study can be extended to apply to automatic categorization problems in a variety of settings.

18. Confidentiality

Describe the procedures you will use to maintain the confidentiality of any personally identifiable data.

All paperwork resulting from this research will only be handled by the people assigned to work on this research, which are: Dr. Nicholas Evangelopoulos, and Ms. Leticia Anaya. The DSCI 4520, 5240, and 5320 course instructors have agreed to award participants with extra credit points and would like to have the names of the students who would have completed the survey. The survey has been adjusted to accommodate the gathering of these names. However, as soon as the instructors get the lists of students who completed the survey, the survey responses will be coded and the names will be deleted to preserve anonymity. Subsequently, survey responses and names of participants will be kept in separate places. All instruments will be maintained under lock and key for a period of three years. The survey responses will be analyzed using published text mining and statistical techniques. No names will be posted on coded surveys, which will be administered in a way that maintains confidentiality of individual information at all times. Outcomes of this study that may be published or presented will only include aggregate results and statistical summaries.

Please specify where your research records will be maintained, any coding or other steps you will take to separate participants' names/identities from research data, and how long you will retain personally identifiable data in your research records. Federal IRB regulations require that the investigator's research records be maintained for 3 years following the end of the study.

The names will be separated from the corresponding survey answers at an early stage. Subsequently, Ms. Leticia Anaya will code survey responses with the use of random re-ordering and system-generated response ID numbers, so that anonymity of research data is maintained. Research records will be maintained under lock and key for a period of three years in the researchers' offices.

19. Publication of Results

Please identify all methods in which you may publicly disseminate the results of your study.

- | | |
|--|---|
| <input checked="" type="checkbox"/> Academic journal | <input checked="" type="checkbox"/> A thesis or dissertation for one of your students |
| <input checked="" type="checkbox"/> Academic conference paper or public poster session | <input type="checkbox"/> Other - Please list below.
(E.g. website or blog) |
| <input type="checkbox"/> Book or chapter | |

Investigator Signature

I certify that the information in this application is complete and accurate. I agree to conduct this study in accordance with the UNT IRB Guidelines and the study procedures and forms approved by the UNT IRB. I understand that I cannot initiate any contact with potential human subjects until I have received written UNT IRB approval.



Nicholas
Evangelopoulos

11/30/2010

Signature of Investigator

Date

Signature of Co- Investigator

Date

Electronic Submission Checklist

1. Print and sign this page and then scan the signed document.
2. Attach all supplementary documents, including:
 - a. Copies of all NIH Training completion certificates not previously submitted to the IRB Office;
 - b. A copy of any proposal for internal or external funding for this study;
 - c. A copy of all recruitment materials;
 - d. A copy of all informed consent forms; and
 - e. A copy of all data collection instruments, interview scripts and intervention protocols.
3. E-mail the application (including this Signature Page) and all supplementary documents to untirb@unt.edu. Please insert "Expedited or Full Board Review" in the subject line of your email.

Contact Shelia Bourns at Shelia.Bourns@unt.edu for any questions about completion of your application.

APPENDIX H
LDA ALGORITHM INFORMATION

The Steyvers and Griffiths Latent Dirichlet Algorithm (LDA) (2004) that uses the Markov Chain Monte Carlo Process (Gibbs Sampling Method) can be found in the following website:

http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

This link was published in: Steyvers, M., and Griffith, “Probabilistic Topic Model,” Chapter in T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum (2007).

For a probabilistic topic model, the Bayesian formula to determine the probability of finding a word in a document was derived by Hoffman (1999) and is given by:

$$P(w_i) = \sum_{j=1}^T P(w_i/z_i = j)P(z_i = j) \quad (9)$$

where $P(Z_{i=j})$ =Probability that the topic= j for the i th word and $P(W_i/Z_{i=j})$ =conditional probability that of the word w_i given the fact that topic j was selected for the i th word. The main difference between pLSA and LDA is that the key mixture probabilities now follow the Dirichlet Distribution which is given by:

$$Dir(\alpha_1, \dots, \alpha_T) = \frac{\Gamma \sum_j \alpha_j}{\prod_j \Gamma \alpha_j} \prod_{j=1}^T P_j^{\alpha_j - 1} \quad (7)$$

Blei et al.(2003) used the Dirichlet Multinomial distribution in determining the probability of finding a word given parameters (α and β) and discovered that the equation that developed was an intractable equation and this is given below:

$$p(w|\alpha, \beta) = \frac{\Gamma \sum_i \alpha_i}{\prod_i \Gamma \alpha_i} \int (\prod_{i=1}^k \theta_i^{\alpha_i - 1}) \left(\prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{i,j})^{w_n^j} \right) d\theta \quad (12)$$

Thus, because it is impossible to compute these probabilities directly when prior and posterior probabilities that follow the Dirichlet distributions, different approaches have been used to estimate these key $p(W/\alpha, \beta)$ probabilities.

The main issue in text mining has been how to determine the probability of finding a word in a document $p(W/\alpha, \beta)$. With LDA, the main issue is how to determine the probability of a word conditioned on the fact that $P(z)$ and $P(w/z)$ are multinomial probabilities from topic mixtures whose probabilities follow *Dirichlet*(α) distributions and from probability distribution of words in a given topic that follow the *Dirichlet*(β) distribution. Bayes' theorem and these probabilities are used to compute the probability of a word $P(W/\alpha, \beta)$.

One of the approaches that are currently being used to solve the untractable probabilities of the Latent Dirichlet Allocation method, is the Steyvers and Griffins(2004) approach which consists of using a Markov Chain Monte Carlo procedure called Gibbs Sampling.

Hogg, et al., (2005) explains the procedure for Gibbs Sampling general procedure. To start out with, a stream of X_0 values are initialized at time $t=t_0$. Then a conditioned random variable Y_i/X_{i-1} is generated from a distribution $f_{y/x}(y/x)$. This conditioned Y_i values are then substituted into $f_{x/y}(x/Y)$ distribution to generate a new set of conditioned X_i values or $X_i/Y_i \sim f_{x/y}(x/Y)$ and the process repeats itself. What is noted here is that the new state of the system only depends on the previous state and not on the past history. Also, the movement from one state to another is on a random basis. These two concepts are the basis for this Gibbs Sampling procedure being called a Markov Chain Monte Carlo—the future state only depends on the present state and not on its history and moving from one state to another state occurs on a random basis.

In this Gibbs Sampling approach, Steyvers and Griffiths (2004) decided to define the posterior distribution of $P(z/w)$ as being proportional to the product of an estimate of $\varphi^{(j)} = P(w/z=j)$ and to an estimate of $\theta^{(d)} = P(z)$. In Griffiths and Steyvers's *Probabilistic Models* (2007) also defined the equation to select the next topic for a word token w_i given previous topics have been chosen for previous words as being proportional to these different terms and this equation is given below.

$$P(z_i = j | z_{-i}, w_i, d_i) \propto \frac{c_{w,j}^{WT} + \beta}{\sum_{w=1}^W c_{w,j}^{WT} + W\beta} \frac{c_{d,j}^{DT} + \alpha}{\sum_{t=1}^T c_{d,j}^{DT} + T\alpha} \quad (13)$$

The first term represents an estimate of $\varphi^{(j)} = P(w/z=j)$ and the second term represents an estimate of $\theta^{(d)} = P(z)$. A suggested good value for $\alpha = 50/T$, where T =number of topics and $\beta=0.01$. In this formula, C^{WT} and C^{DT} are matrixes with dimensions W by T and D by T respectively. C^{WT}_{wj} contains the number of times word w is assigned to topic j , not including the current instant i and C^{DT}_{dj} contains the number of times topic j is assigned to some work token in document d , not including the current instance i . The distributions $P(Z_{i=j})$ and $P(W_i/Z_{i=j})$ can be obtained from the C^{WT} and C^{DT} matrixes, either after each C^{WT} and C^{DT} is being computed or after a specific number of iterations have updated the C^{WT} and C^{DT} matrixes. The estimate of the probabilities $\varphi^{(j)} = P(w / z=j)$ and $\theta^{(d)} = P(z)$ used to compute the probability of finding a word w_i are given by:

$$\Phi_i^{(j)} = \frac{c_{w,j}^{WT} + \beta}{\sum_{w=1}^W c_{w,j}^{WT} + W\beta} \quad (14)$$

$$\theta_j^{(d)} = \frac{c_{d,j}^{DT} + \alpha}{\sum_{j=1}^T c_{d,j}^{DT} + T\alpha} \quad (15)$$

How the above equations follow the Gibbs Sampling procedure is briefly outlined here. At time $t=t_o$, the topics are selected randomly for a word w_i in a document d_j . From these initial

values the C^{WT} and C^{DT} matrices are computed. These matrices are then updated to compute the terms. $\varphi^{(j)} = P(w / z=j)$ and $\theta^{(d)} = P(z)$. Then these terms are used to compute the probability of selecting the next topic for w_i in a document d_j or $P(z_i=j/z_{-i}, w_i, d_i, \dots)$. Once the topics have been updated for each word w_i , new matrices C^{WT} and C^{DT} are computed and the process keeps going until it converges and there is not much difference in the matrices after a number of iterations. This following flowcharts shows the basic algorithm layout.

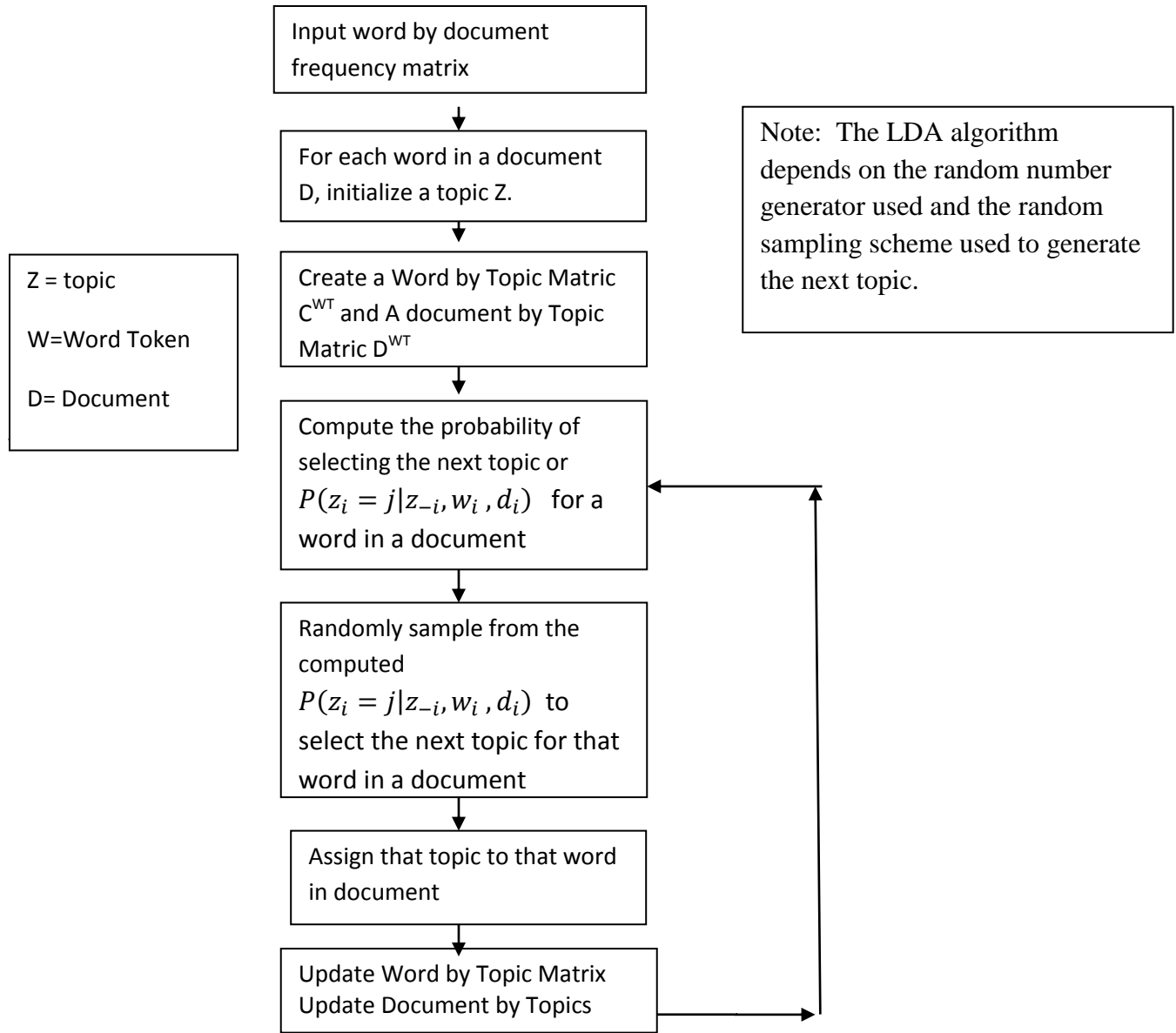


Figure H.1. Basic LDA flowchart © Leticia H. Anaya.

REFERENCES

- Abramson, N., Keating, R., and Lane, H. (1996). Cross national cognitive process differences: A comparison of Canadian, American and Japanese managers. *Management International Review*, 36(2), 123-147.
- Ahm, W., Kim, N. S., Lassaline, M. E., and Dennis, M.J. (2000). Casual status as a determinate of feature centrality. *Cognitive Psychology*, 41, 361-416.
- Anonymous. (2011). On a mission to eradicate paper. *Health Management Technology*, 32(3), 16-17.
- Arterberry, M. E., and Bornstein, M. H. (2002). Infant perceptual and conceptual categorization: The roles of static and dynamic stimulus attributes. *Cognition*, 86, 1-24.
- Ashby, F.G., and Alfonso-Reese, I.A. (1995). Categorization as probability density estimators. *Journal of Mathematical Psychology*, 39, 216-233.
- Ashby F.G, Gott R.E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learn, Memory and Cognition*, 14, 33-53.
- Ashby, F. G. and Maddox, T.W. (2005) Human category learning. *Annual Review Psychology*, 56, 149-78.
- Ashby F.G, Townsend JT. (1986). Varieties of perceptual independence. *Psychology Review*, 93, 154-179.
- Ashby F.G and Maddox W.T. (1998). Stimulus categorization. In M.H. Birnbaum (Ed.), *Handbook of perception and cognition: Measurement, judgment, and decision making* (pp. 251- 301). New York: Academic.
- Ashby F.G., and Waldron E.M. (1999). On the nature of implicit categorization. *Psychology Bulletin Review*, 6, 363-378.
- Ashby, F. G. (2001). Categorization and similarity models: Neuroscience applications. In *International encyclopedia of the social and behavioral sciences* (pp. 1535-1538). Amsterdam: Pergamon Press.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Bornstein, M. H. (1984). A descriptive taxonomy of psychological categories used by infants. In C. Sophian (Ed.), *Origins of cognitive skills*, (pp. 313-338). Hillsdale, N.J. Erlbaum.

- Brooks L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch, BB Lloyd (Ed) *Cognition and categorization*, (pp. 169–211). Hillsdale, NJ: Erlbaum.
- Bruce, R. (1990). Simple tools solve complex problems. *Quality*, 29(4), 50-51.
- Cios, K. J., Pedrycz, W., Swiniarski, R. W., and Kurgan, L. A. (2007). *Data mining: A knowledge discovery approach*. Springer Science-Business Media, LLC.
- Cruz, B. V. (1998). Against the explicit/implicit distinction. *Revista Alicantina de Estudios Ingleses*, 11, 241-258.
- Day, R. L., and Bodur, M. (1978). Consumer response to dissatisfaction with services and intangibles. *Advances in Consumer Research*, 5(1), 263-272.
- De Chernatony, L. and Harris, F. (2000). *The challenge of financial services branding: Majoring on category or brand values?* Buckingham: Open University.
- Estes W.K. (1986). Array models for category learning. *Cognitive Psychology*, 18, 500-549.
- Estes W.K. (1994). *Classification and cognition*. New York: Oxford Univ. Press.
- Evans, J. R., and Lindsay, W. M. (1999). *The management and control of quality* (4rth ed.). Cincinnati, OH: International Thomson Publishing.
- Eylon, D., and Au, K.,Y. (1999). Exploring empowerment cross-cultural differences along the power distance dimension. *International Journal of Intercultural Relations*, 23(3), 373-385.
- Feldman, R., and Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. New York, NY: Cambridge University Press.
- Ferguson, J. L., and Johnston, W. J. (2011). Customer response to dissatisfaction: A synthesis of literature and conceptual framework. *Industrial Marketing Management*, 40, 118-127.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.
- Gluck, M. A., & Bower, G. H. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, 27, 166-195.
- Gluck, M. A., & Bower, G. H. (1988b). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117, 227-247.
- Gopal, S., and Yang, Y. (2010). Multilabel classification with meta-level features. *Proceeding of the 33rd international ACM SIGIR conference on Research and development in Information retrieval (SIGIR '10)*, (pp. 315-322). New York, NY: ACM.

- Griffiths, T.L., Sanborn, A.N., Canini, K. R., and Navarro, D. J. (2008). Categorization as nonparametric bayesian density estimation. *The probabilistic mind: Prospects for Bayesian cognitive science*. New York: Oxford University Press Inc.
- Griffiths T. L., and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science of the United States of America*, 101, 5228-5235.
- Griffiths, T. L., Steyvers, M., and Firl, A. (2007). Google and the mind. *Psychological Science*, 18, 1069-1076.
- Gupta, S., McLaughlin, E., and Gomez, M. (2007). Guest satisfaction and restaurant performance. *Cornell Hotel and Restaurant Administration Quarterly*, 48(3), 284-298.
- Gwet, K. L. (2008). Variance estimation of nominal-scale inter-rater reliability with random selection of raters, “ *Psychometrika*, 73(3), 407-430.
- Han, J., and Kamber, M., (2006). In Morgan Kaufmann Publishers (2nd Ed.), *Data mining: concepts and techniques*. San Francisco, CA: Elsevier.
- Hintzman D.L. (1986). Schema abstraction in a multiple-trace memory model. *Psychology Review*, 93, 411–428.
- Hirschman, E. C. and Krishnan, S. (1981). Subjective and objective criteria in consumer choice: An examination of retail patronage criteria. *Journal of Consumer Affairs*, 15(1), 115-127.
- Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177-196.
- Hofmann, T. (2000). Learning the similarity of documents : an information-geometric approach to document retrieval and categorization. *Advances in Neural Information Processing Systems* (pp. 914-920). Boston, MS: MIT Press.
- Hogg, R., McKean, J., and Craig, A. (2005). *Introduction to mathematical statistics*, (6th ed.). New River, NJ: Pearson: Prentice Hall.
- Homa D, Sterling S, Trepel L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology : Learn, Memory, and Cognition*, 7, 418–439.
- Heuer, M., Cummings, J., and Hutabarat, W. (1999). Cultural stability or change among managers in Indonesia? *Journal of International Business Studies*, 30(3/3), 599-610.
- Iglesias, V., (2004). Preconceptions about service: How much do they influence quality evaluations? *Journal of Service Research*, 7(1), 90-102.

- Jamain, A., and Hand, D.(2008). Mining supervised classification performance studies: A meta-analytic investigation. *Journal of Classification*, 25, 87-112.
- Janowsky J.S., Shimamura A.P., Kritchevsky M., and Squire L.R. (1989). Cognitive impairment following frontal lobe damage and its relevance to human amnesia. *Behavioral Neuroscience*. 103, 548–560.
- Juran, J. M., and Gryna, F. M. (1993). *Quality planning and analysis* (3rd. ed.). New York, NY: McGraw-Hill, Inc.
- Klimt, B., and Yang, Y. (2004) The Enron corpus: A new dataset for e-mail classification research. *European Conference on Machine Learning, September 20-24, 2004, Pisa, Italy*.
- Knowlton B. J., Mangels J. A., Squire L. R. (1996a). A neostriatal habit learning system in humans.” *Science*, 273, 1399-1402
- Knowlton, B. J., Squire, L. R., Paulsen, J. S., Swerdlow, N. R., Swenson, M., Butters, N. (1996b). Dissociations within nondeclarative memory in Huntington's disease. *Neuropsychology*, 10, 538-548.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 22-24.
- Kruschke, J. K. (1996). The role of base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(1), 3-26.
- Kruschke, J. K. (2001). The inverse base-rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1385-1400.
- Kruschke, J. K. (2003). Attentional theory is a viable explanation of the inverse base rate effect: A reply to Winman, Wennerholm, and Juslin (2003). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1396-1400.
- Lamberts K. (2000). Information-accumulation theory of speeded categorization. *Psychology Review*, 107, 227-260.
- Landis, R., and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. 33(1), 159-174.
- Landauer, T.K., and Dumais, S. T. (1987). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 221-240.
- Lavengood, K., and Kiser, P. (2007).“Information professionals in the text mine. *Online*, 31(3), 16-21.

- Lee, S., Song, J., and Kim, Y. (2010). An empirical comparison of four text mining methods. *Journal of Computer Information Systems, Fall 2010*.
- Leng N.R. and Parkin A.J. (1988). Double dissociation of frontal dysfunction in organic amnesia. *British Journal of Clinical Psychology*, 27, 359–362.
- Lewandowsky S, Kalish M, Ngang SK. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General*. 131, 163-193.
- Lewis, R. C., and Booms. B. H. (1983). The marketing aspects of service quality. In L. Berry, G. Shostack, and G. Upah, (Ed), *Emerging perspectives on services marketing*, (pp. 99-107). Chicago: American Marketing.
- Li F., Yang Y. (2003). A loss function analysis for classification methods in text categorization. *International Conference on Machine Learning (ICML)*, 472-479.
- Maddox W.T., Ashby F.G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioral Processes*, 66, 309-332.
- Markman AB, Ross BH. (2003). Category use and category learning. *Psychology Bulletin*, 129, 592–613.
- Medin D.L., Schaffer M. M. (1978). Context theory of classification learning. *Psychology Review*, 85, 207–238.
- Medin, D.L. and Edelson, S.M. (1988). Problem structure and the use of base rate information from experience. *Journal of Experimental Psychology: General*, 117, 68-85.
- Mosley, R. (2007). Customer experience, organizational culture and the employer brand. *Brand Management*, 15(2), 123-134.
- Namata et al. (2009). Collective classification for text classification. *Text mining: Classification, clustering, and applications*, Taylor and Francis Group, LLC.
- Nosofsky R.M. (1986). Attention, similarity, and the identification categorization relationship. *Journal of Experimental Psychology : General*, 115, 39–57.
- Papadimitriou, C.H., Raghavan, P., Tamaki, H., and Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61, 217-235.
- Ogure, T., Nakabo, Y., Jeong, S., and Yamada, Y. (2009). Hazard analysis of an industrial upper-body humanoid. *Industrial Robot: An International Journal*, 36(5), 469-476.
- Pantelidis, I. S. (2010). Electronic meal experience: A content analysis of online restaurant comments. *Cornell Hospitality Quarterly*, 51, 483.

- Parasuraman, A., Zeithaml, V. A., and Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, 4(4), 41-50.
- Pinkster, R. (2008). An empirical examination of competing theories to explain continuous disclosure technology adoption intentions using XBRL as the example technology. *The International Journal of Digital Accounting Research*, 8(4), 81-96.
- Posner M.I., and Keele S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77, 353–363.
- Posner M.I., and Keele S.W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, 83, 304–308.
- Ravinsankar, P., Ravi, V., Rao, G. R., and Bose, L. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, 50, 491-500.
- Rehder, B. (2003). A casual-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1141-1159.
- Retrieved from Google <http://www.hss.iitb.ac.in/courses/HS435/Neurolinguistics.htm> accessed on August 21, 2011
- Retrieved from Google <http://inside-the-brain.com/2011/02/09/pioneering-brain-surgery-to-treat-tourettes/> accessed on August 21, 2011.
- Retrieved from Facebook <http://www.facebook.com/press/info.php?statistics>, accessed on August 21, 2011.
- Retrieved from LinkedIn http://www.linkedin.com/network?trk=hb_tab_net, accessed on August 21, 2011.
- Rodgers , S. (1998). Sizing up the Internet. *Credit Union Magazine*, 64(7), 8.
- Salton, G. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.
- Shmueli, G., Patel, N.R., and Bruce, P.C. (2007). *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*, Hoboken, NJ: John Wiley & Sons, Inc.
- Smith J.D., and Minda J.P. (1998). Prototypes in the mist: the early epochs of category learning. *Journal of Experimental Psychology: Learn, Memory, and Cognition*, 24, 1411-1430.

- Sperber, D., and Wilson, D. (1986/1995). *Relevance: Communication and Cognition*, 2nd Edition. Cambridge, MA: Harvard University Press.
- Spira, J. (2006). Lost knowledge: File not found. *KM World*, 15(10), 1.
- Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.
- Steyvers, M., and Griffiths, T. L., (2008). Rational analysis as a link between human memory and information retrieval. In N. Chater and M. Oaksford (Ed.), *The probabilistic mind: Prospects for Bayesian cognitive science*. New York: Oxford University Press.
- Takeuchi, H. and Quelch, J. A. (1983). Quality is more than making a good product. *Harvard Business Review*, 61(4), 139-145.
- Tam, V., Santoso, A., and Setiono, R., (2002). A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization. *International Conference on Pattern Recognition*, 4, 40235.
- Tang, L., Rajan, S., and Narayanan, V. (2009). Large scale multi-label classification via metaLabeler. *International World Wide Web Conference Committee (IW3C2)*, April 20-24, 2009, Madrid, Spain, 211-220.
- Taner, M. T., and Sezen, B., (2007). An overview of six sigma applications in healthcare industry. *International Journal of Health Care Quality Assurance*, 20(4), 329-340.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), 1566-1581.
- Vanpaemel, W., Storms, G., and Ons, B. (2005). A varying abstraction model for categorization. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Erlbaum.
- Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.
- Vercellis, C. (2009). *Business intelligence: Data mining and optimization for decision making*. Hoboken, NJ: John Wiley and Sons, Inc.
- Wall, D. (2001). *Crime and the Internet: Cyber crimes and cyber fears*. New York, NY: Routledge.
- Weiss, S. M., Indurkha, N., Zhang, T., and Damerou, F. J. (2005). *Text mining: Predictive methods for analyzing unstructured information*. Springer Science-Business Media, Inc.

Winman, A., Wennerholm, P., and Juslin, P. (2003). Can attentional theory explain the inverse base rate effect? Comment on Kruschke (2001). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1390-1395.

Willis, A. J., Nouri, M., and Moberly N. (2006). Formation of category representations. *Memory and Cognition*, 34(1), 17-27.

Yang Y. and Liu X. (1999). A re-examination of text categorization methods. *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 42-49.

Yeh, J., Wu., T., and Tsao, C. (2011). Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems*, 50, 439-448.