12-2016

Towards building a review recommendation system that trains novices by leveraging the actions of experts

Shilpa Khanal
University of Nebraska-Lincoln, shilpa.khanal@gmail.com

Towards building a review recommendation system that trains novices

by leveraging the actions of experts

By

Shilpa Khanal

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Leen-Kiat Soh

Lincoln, Nebraska

December, 2016

Towards building a review recommendation system that trains novices by leveraging the actions of experts

Shilpa Khanal, M.S.

University of Nebraska, 2016

Advisor: Leen-Kiat Soh

Online reviews increase consumer visits, increase the time spent on the website, and create a sense of community among the frequent shoppers. Because of the importance of online reviews, online retailers such as Amazon.com and eOpinions provide detailed guidelines for writing reviews. However, though these guidelines provide instructions on how to write reviews, reviewers are not provided instructions for writing product-specific reviews. As a result, poorly-written reviews are abound and a customer may need to scroll through a large number of reviews, which could be up to 6000 pixels down from the top of the page, in order to find helpful information about a product (Porter, 2010). Thus, there is a need to train reviewers to write better reviews, which could in turn better serve customers, vendors, and online e-stores. In this Thesis, we propose a review recommendation framework to train reviewers to better write about their experiences with a product by leveraging the behaviors of expert reviewers who are good at writing helpful reviews.

First, we use clustering to model reviewers into different classes that reflect different skill levels to write a quality review such as expert, novice, etc. Through temporal analysis of reviewer behavior, we have found that reviewers evolve over time, with their reviews becoming better or worse in quality and more or less in quantity. We also investigate how reviews are valued differently across different product categories. Through machine learning-based classification techniques, we have found that, for

products associated with prevention consumption goal, *longer reviews are perceived to be more helpful*; and, for products associated with promotion consumption goal, *positive reviews are more helpful than negative ones*.

In this Thesis, our proposed review recommendation framework is aimed to help a novice or conscientious reviewer become an expert reviewer. Our assumption is that a reviewer will reach the highest level of expertise by learning from the experiences of his or her closest experts who have a similar evolutionary pattern to that of the reviewer who is being trained. In order to provide assistance with intermediate steps for the reviewer to grow from his or her current state to the highest level of expertise, we want to recommend the positive actions—that are not too far out of reach of the reviewer—and discourage the negative actions—that are within reach of the reviewer—of the reviewer's closest experts. Recommendations are personalized to fit the expertise level of reviewers, their evolution trend, and product category. Using the proposed review recommendation system framework we have found that for a random reviewer, at least 80% of the reviews posted by closest experts were of higher quality than that of the novice reviewer. This is verified in a dataset of 2.3 million reviewers, whose reviews cover products from nine different product categories such as Books, Electronics, Cellphones and accessories, Grocery and gourmet food, Office product, Health and personal care, Baby, Beauty, and Pet supplies.

Acknowledgements

I would like to thank my advisor, Dr. Leen-Kiat Soh for his constant support and guidance from the inception to the completion of this research. I also would like to thank my committee members, Dr. Ashok Samal and Dr. Peter Revesz for their valuable time.

Lastly, I'd like to dedicate this thesis to my parents and my husband who have been a source of encouragement and love for me throughout my life.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

When consumers shop online, they go through product information to evaluate different products, and have access to number of product reviews (Mudambi and Schuff, 2010). Online consumer reviews are provided in addition to product description and share consumers' personal usage experiences with the product. The buyer-created review information compliments the seller-created product attribute information making online consumer reviews work effectively as sales assistant to help consumers identify the products that best match their usage conditions (Chen and Xie, 2008).

Online consumer product review is an emerging electronic market phenomenon which is playing an important role in deciding consumers' purchase behavior (Chen and Xie, 2008). The existence of consumer reviews on a website has proved to increase the usefulness and social presence of the website (Kumar and Benbasat, 2006). Online reviews tend to increase consumer visits, increase the time spent on the website, and create a sense of community among the frequent shoppers. Because of the importance of online reviews, online retailers such as Amazon.com and eOpinions even go as far as posting detailed guidelines for writing reviews as helpful consumers reviews are perceived to be highly valuable (Mudambi and Schuff, 2010).

However, though these guidelines provide instructions on how to write reviews, reviewers are not provided instructions for writing product-specific reviews—that is,

what kind of information is perceived as more helpful for which product. Furthermore, it may be costly to treat reviewers as writers who need practices and time to become good at writing reviews.

1.1 Background

Meanwhile, for a very long time *recommendation systems* have been providing recommendations on web pages, videos, movies, music and books (McDonald *et al*., 2000). Recommendation systems have been helping customers to purchase products from E-commerce website (Schafer *et al*., 2001). High volume of digital data has increased the need of recommendation systems on different types of digital content.

On the basis of what is recommended, there are different types of recommendation system found in web today- 1) *expertise recommenders*, 2) *item recommenders*, and 3) *action recommenders*. *Expertise recommender* (McDonald *et al*., 2000) is a system that helps users to locate experts in domain specific task and make referrals based on the users' expertise rating. The system has been tested to recommend experts in Tech Support system. Similarly, Expert Finder (Vivacqua *et al*., 2000) is an agent that classifies users as experts and novices by analyzing documents created by them in their day-to-day life. It has been used to segregate Java Programmers and assign numerical value to define their level of expertise. Novice users can then ask questions to experts. *Item recommenders*, as the name implies, recommend items to users. These are the most popular type of recommender systems especially used by e-commerce websites. These systems analyze users' purchase history and recommend items that are most similar to previously bought items. *Action recommenders* are the recommendation systems that produce suggestions or advices to customers in the form of actions or plans.

The recommended actions or plans are varied with types of customers. 'Role Models' approach (Yang *et al*., 2002) is used in which failed customers are recommended to perform similar actions as active customers or role models. For example, in banking system, for the rejected loan applicants the system recommends certain actions that will increase their chances of receiving loan in next attempt. The recommended actions will move the rejected applicant more closely to the accepted applicant.

In this thesis, we focus on *action recommenders* for recommending actions to improve buyers ability to write helpful reviews. For this, we first find role models. These are reviewers who have the skills to write most helpful reviews. Before going into the problem of finding these so-called role models, we need to look at *user modeling*.

Computer user modeling is the process of gathering information about users, and using the information to adapt the underlying system to the users' needs (Kobsa, 2001). User modeling caters to individual user needs and interests, which is especially useful for complex, widely available software (Fisher, 2001). User modeling has been an important part of recommendation systems to suggest products to users and is widely used in commercial websites like Netflix, Amazon.com and so on (Jameson, 2009). These websites infer user interests or taste on products and make suggestions accordingly. User modeling stores information about individual users and uses this information to assist the system in "serving the user better" (Biswas and Robinson, 2010).

For example, one traditional application of user modeling is scheduling meetings or appointments (Gervasio *et al.*, 2005; Jameson, 2009). The system assists in the task of entering these meeting schedules in users' calendar, based on the users' preferences on meeting types, times and locations. The main goal of these systems is to learn about users

and help them save time and effort to perform daily tasks. Another similar example application of user modeling is a multimedia conversation system that helps to search real state based on users' personal and financial preferences (Zhou & Aggrawal, 2004). The conversation between system and users help to gather information regarding users' preferences on real state. These diverse user queries are used to generate response tailored for the specific user. In both these examples, the system first models users from the past or current interactions with the system and then serves the users based on their tastes or preferences.

In this thesis, we use user modeling to find and understand different models of reviewers based on their characteristics. We then interpret each model based on their expertise level measured by the perceived helpfulness of their reviews. Further, each type of reviewers is further analyzed to understand how their expertise evolve over years. With the knowledge of expertise level and their evolution, we develop a framework for recommending actions to help novice reviewers write better reviews by performing similar actions as expert reviewers.

1.2 Motivation

The main goal of our thesis is to perform user analysis on reviewers to understand their behaviors and how those behaviors change with time, and develop a framework for a recommendation system for helping reviewers. We aim to explore the possibility of training reviewers in a cost-effective way to better write about their experiences with the product by leveraging the behaviors of expert reviewers who are good at writing helpful reviews. Training reviewers to write better reviews is important in two ways. First, better reviewers write better reviews and thus it would reduce the number of bad reviews.

Second, reviewers receiving training could become better reviewers faster than someone without the training.

Reducing the number of bad reviews is beneficial to all involved parties-customers, vendors, and Amazon.com. Current state-of-art suggests that, for example, on Amazon.com, to find product information customers have to scroll through a large number of reviews which could be up to 6000 pixels down from the top of the page (Porter, 2010). This suggests that customers are required to invest more time to search reviews that provide required product information and help them make purchase decision. Hence there is a need for Amazon.com to discourage bad reviews and encourage good reviews in order to save customers' product search time and provide better *user experiences*.

Further, to facilitate customers find better reviews quickly and with ease, Amazon has been running its Vine program since 2007. This program invites trusted customers to become Vine voices based on their reviewer rank, which is a reflection of the quality and helpfulness of their reviews as judged by other Amazon customers (Puranam *et al*., 2014). Amazon.com promotes the reviews written by Vine members by posting them on the top of reviews chart and encouraging other customers to read them. Although Vine program *passively* trains customers to review their own usage experiences by exposing them to Vine voices who are the best reviewers of Amazon.com, it doesn't provide insights on actions that can lead to quality review. An illustration by examples of step-by-step actions that novices should adhere to in order to write better reviews can speed up this training process.

Training novice reviewers to learn to write quality reviews *faster* is beneficial to vendors who use Amazon.com as a platform to launch new products. When new products are introduced, the number of reviews are few, and the impact of these initial reviews are even more significant. More specifically, the impact of online reviews on its sales is maximum when the product is new and the impact decreases as the product ages over time (Hu *et al*., 2008). Thus, it is sub-optimal to have novices review new products as they may not provide the product information or even worse they may write bad reviews without explaining it in details. The sooner the reviewers are trained to write quality reviews, the better it is for the sales of new products. In addition, when launching new products on behalf of participating vendors, Amazon.com provides Vine members with free products submitted to the program by those vendors (Puranam *et al*., 2014). However, there is a limited number of Vine members. Hence it may be risky for vendors to rely on just those Vine members to provide reviews on new products as their products might go to the wrong reviewers. Therefore it becomes more important to train novice reviewers *faster* to write quality reviews specially when they are expected to review a new product.

Therefore, there is a need to train novice reviewers to write better reviews for better serving customers, vendors, and online e-stores such as Amazon.com but it hasn't been addressed so far. Current state-of-the-art suggests that reviewers write product reviews based on their own ability to articulate their experiences with the product (Dellarocas *et al*., 2010). Amazon.com, for example, does not manage reviewers in any tangible way that is to say reviewers are acting on their own ability (Porter, 2010). It may be prohibitively costly to treat each reviewer as a writer and train them with proper

review-writing skills. Implementation of the system that we envision in this Thesis is a cost effective approach to train reviewers so they do not have to rely on their individual judgment on what product information to share or how to present product information for writing helpful reviews. Implementation of our system incurs no monetary cost on the part of customers or vendors. The problems of developing this framework is detailed in Section 1.3.

1.3 Problem statement

In this thesis, we focus on two sub-problems of developing a framework to provide action recommendations to reviewers to help them write quality reviews. The first sub-problem is to differentiate reviewers based on the *quality* of their reviews to identify, for example, reviewers with high quality reviews as *expert* reviewers and those with low quality reviews as *novice* reviewers. The second sub-problem is to devise an approach to leverage expert reviewers' behaviors to help train novice reviewers effectively and efficiently.

The first sub-problem of distinguishing reviewers into different classes is based on their ability to write quality reviews. There are many problems within this sub-problem such as: (1) defining review quality, (2) defining review quality based on product type, (3) distinguishing expert reviewers from other reviewers, and (4) finding different classes of reviewers. The review quality is a reflection of review credibility and persuasiveness. The persuasiveness of a review largely depends on how users perceive online reviews. Persuasiveness of review can be discussed from two sides (1) the retailer perspective and (2) customer perspective. From the **retailer perspective**, a review is considered of good quality if it increases product sales by convincing customers to

purchase the product (Chevalier and Mayzlin, 2006; Lee *et al*., 2008). From the **customer perspective**, a review is considered of good quality if it helps them to make an informed decision which may or may not lead to the purchasing of product. From both retailer and customer perspective, review quality is subjective and therefore difficult to quantify. The problem of quantification of review quality is further aggravated when product type is considered. The customer perception of review quality may be different based on product types as customers tend to search for different information based on product types. For example, customers look for different information when searching products that are *readily* available compared to the products *scarcely* available in the market (Dellarocas *et al*., 2010). Also, based on the pre-consumption goal of customers, the expectations from reviews could be different (Nelson, 1970). Therefore reviews should serve different purpose depending on the product type such as hit or niche, search or experienced, etc. In this thesis, we try to solve the problem of categorization of products into different types and examine salient review features that improve the review quality for different product types. Moreover, it is difficult to distinguish expert reviewers from other classes of reviewers. Should the expertise of reviewers be measured entirely from the review quality or other features such as reviewing frequency, active period, product type and so on? Given the large number of reviewers and even larger number of reviews, the task of differentiating reviewers into different classes becomes complex. Also, how many classes of reviewers should be labeled? How do we come up with the threshold that distinguishes one class from another? This Thesis tries to answer these questions.

The second sub-problem that we will investigate in this Thesis is devising an approach to leverage expert reviewers' behaviors to help train novice reviewers effectively and efficiently. There are many issues within this sub-problem such as- (1) defining expert reviewers' behavior, (2) choosing the reviewers class whom to make these recommendations, (3) developing ways to make recommendations, and (4) making recommendation process effective and efficient. Reviewers perform different actions that make them experts ranging from writing few high quality review consistently, or writing many but both high and low quality reviews that would average to high quality review. There seems to be no pre-defined course of actions which makes a reviewer expert or novice. So the task of defining experts' actions requires addressing these issues. Furthermore, the task of making recommendations effectively and efficiently requires answering two other questions: (1) which reviewer should be recommended or in other words, how do we find reviewer who requires training, and (2) what kind of actions should be recommended. The task of choosing reviewers to direct recommendations to is tricky as we need to answer questions such as "do we recommend actions to someone who is motivated to review and lacks reviewing skills?" or "do we recommend actions to someone who is not interested and lacks reviewing skills?". We need to develop strategies to first differentiate reviewers' motivation level and then prioritize one over another. Additionally, recommended actions might be different based on both reviewers' motivation and skill level as well as product types. Intuitively, the actions recommended to new and inexperienced reviewer should be more basic and elaborate than old and experienced reviewer. Additionally, recommended actions could be product type specific i.e., certain kind of information is perceived as more helpful for certain products. This

Thesis tries to provide solutions to the problems of understanding the skill, experience and motivation level of reviewer and then device appropriate actions to recommend in order to help them review better .

1.4 Solution Approach

In this section we discuss the solutions of the two sub-problems: (1) differentiating reviewers based on the *quality* of their reviews to identify different classes of reviewers, and (2) devising an approach to leverage expert reviewers' behaviors to help train novice reviewers effectively and efficiently.

As discussed in Section 1.2, one of the motivations of our research is to train reviewers to write quality reviews in order to facilitate customers find good reviews quickly and with ease. We focus on customers perception of review quality. Customers use online product reviews provided by consumers of the product as a major information source to evaluate the product quality (Hu *et al*., 2008). Amazon.com implements a voting system in which customers rank a review if it helped them to know more about the product and decide on buying the product or not. The proportion of helpful votes reflects how content customers are with the review and if it helped them in making purchase decision (Chen *et al*., 2008; Korfiatis *et al*., 2012). The review quality can be measured by the proportion of helpful votes received by the review (Chen *et al*., 2008). Since the customer perspective of review quality is important to meet our motivation, we use review helpfulness as a measure of review quality. Additionally, to understand how reviews are valued differently across different products, we look into the relationship between review quality measured in terms of helpfulness, with other review features such as its rating score, length, and so on for diverse product categories. This analysis helps to

find salient features of quality review for each product category. Further, based on the degree of review quality we distinguish expert reviewers by using clustering algorithm. The clustering algorithm divides reviewers into optimal number of classes based on their behaviors such as number of reviews they have posted, degree of helpfulness of their reviews, rating score, length of their reviews and their active period as reviewers. Each class is interpreted and labeled based on their motivation and skill level.

We use decision tree to classify reviewers into appropriate classes. The class of a reviewer reflects the motivation and skill level of the reviewer. All classes of reviewers are observed over time and over different product types. The temporal analysis of reviewers shows the evolution of each class of reviewers over time which helps to understand the learning curve of each class of reviewers with respect to their skills for writing good quality reviews. A reviewer belonging to a class that represents low review-writing skill level is chosen to be trained. The reviewer is trained by first choosing his or her closest experts and then using the experts' actions to make appropriate action recommendations. The recommended actions also focus on product specific review writing-style to produce quality reviews by following the action sequence of the closest experts.

1.5 Contributions

This research makes several contributions:

- It creates user models of reviewers with different skill level for posting quality reviews. Each model of reviewers have unique features and exhibit different behaviors. These models can be analyzed by researchers to address a variety of problems related to reviewer behaviors in e-commerce.

- It develops a classifier model that predicts the expertise of a given reviewer. Since the classifiers are product-specific, the prediction of reviewers' expertise is accurate. These classification models can be used by e-commerce websites to assess a reviewer's ability to write quality reviews in real time.
- Reviewers are observed for a length of time to understand their evolution. For some reviewers, evolution may indicate improvement in their reviewing skills whereas for others it may indicate the opposite. These findings give much insights on how to interact with reviewers in the future for generating action recommendations.
- It lays the groundwork for the construction of an action recommendation system for reviewers. The actions performed by an expert reviewer can be recommended to reviewers who have poorer reviewing skills. A real time implementation of this framework could be a highly beneficial for customers and vendors in any ecommerce websites.

1.6 Overview

The remaining part of this Thesis is organized as follows.

Chapter 2 describes related work and background, overviewing related work in the fields of recommendation systems in Section 2.1; user modeling and its applications in Section 2.2; the process of user modeling in Section 2.3 that covers various unsupervised approaches for pattern recognition (in Section 2.3.1) and supervised approaches for validation and interpretation (in Section 2.3.2); and closes with discussion on sentiment analysis in Section 2.4.

Chapter 3 describes methodology. It begins in Section 3.1 describing the method by which reviewers are modeled. This includes defining review quality using various features set and then modeling reviewers into different classes based on their review quality. Section 3.2 talks about building recommendation system framework by analyzing reviewer evolution trends and review sentiment.

Chapter 4 talks about understanding reviewers. It describes how data on which the reviewer models are based was collected in Section 4.1, how the collected data was preprocessed in Section 4.2, how the data clustering was applied to create reviewer models in Section 4.3, and finally how the reviewer-classification for the models were built in Section 4.4.

Chapter 5 proposes recommendation system framework. Section 5.1 presents how different classes of reviewer evolve over time in regard to their reviewing skill. Section 5.2 talks about sentiment analysis highlighting how review sentiment differs from one class of reviewers to another. Section 5.3 details review recommendation system framework that is based on the reviewer evolution to generate review recommendations.

Finally, Chapter 6 draws conclusions from various experiments regarding reviewer classes and their evolution. Section 6.1 focuses on conclusions and Section 6.2 talks about future works.

Chapter 2
Related Work & Background

In this chapter we cover related work in recommendation systems and user modeling. We argue that we apply user-modeling framework to solve a problem not traditionally addressed by the field: the detection and prediction of different classes of reviewers based on their expertise level to write helpful online reviews, and observation of evolution in each class using temporal analysis of the perceived helpfulness of their reviews and using this information to recommend actions to facilitate reviewers write better reviews.

We propose a framework for developing a recommendation system that provides action recommendations to reviewers on how to write better reviews that serves customers to help them make informed purchase decisions. Therefore we cover recommendation systems and their usage patterns. To better understand reviewer behavior, we also utilize machine learning algorithms to create user models, so we describe the basic concepts related to user modeling and its application. We cover various steps of the user modeling process that is crucial for creating user models that accurately represent reviewers. Additionally, we perform opinion mining in product reviews to understand the tone of reviews that are helpful to other customers. In this light, we cover the state-of-art of opinion mining tool and their applications as well in this Chapter.

2.1 Recommendation systems

Recommendation systems have been used to provide recommendations on items such as webpages, videos, movies, music, and books (McDonald *et al*., 2000). Such recommendations are usually personalized in which case the recommended items are different for different users or user groups (Ricci *et al*., 2011). However the recommendations could also be non-personalized which are generally easy to generate and are featured in newspaper or magazines. Amazon.com (Linden *et al*., 2003; Ricci *et al*., 2011) is an e-commerce website which tries to personalize the online store for each customer such as suggesting programming language for software engineers or baby toys for new parents. To generate personalized recommendations for each customer, Amazon.com keeps track of the customer's purchase history and items rated by the customer. There are diverse recommendation algorithms that are applied to generate a list of recommendations. Collaborative filtering (cite) is one of the most popular recommendation algorithm that provides recommendation in a two-step process: (1) it calculates the level of similarity between users based on their rating for common items, and (2) then it predicts the user preference of particular item by calculating the weighted summation of rating provided by the most similar users for that item (Herlocker *et al*., 2004).

Pham *et al*., (2014) talks about an *expert-based* recommendation system which recommends movies to users based on experts' opinions. The recommendation list consists of movies that have high ratings from experts. One of the crucial parts of this system is measuring the expertise level of users to determine the set of experts. The set of experts are different or specific for each user because each user has a different set of

preferences. For example, for a user who prefers action movies in the genre of James Bond, his or her set of experts will have expertise in this particular area.

Recommendation systems are not limited to providing recommendations on items but they also produce suggestions or advices to customers in the form of actions or plans. The recommended actions or plans are varied with types of customers. Service-based corporations try to attract new customers by recommending actions that will result in some kind of benefit for the customers. They try to recommend various actions for failed or low profile customers in order to make them more active or high profiled customers. As stated in Chapter 1, action recommendation systems use 'Role Models' approach (Yang *et al*., 2002) in which failed customers are recommended to perform similar actions as active customers or role models. The recommendation is achieved in three steps (Yang *et al*., 2002). First, data mining techniques are used to find "good" or "positive" customers that are active and are accepted into good class and the "bad" or "negative" customers who are not. Second, from the set of positive customers a number of representative cases of customers that can be used as "role models" for the rest are selected. Third, various actions are recommended to "negative" customers that will switch them into "positive" customers.

Our approach extends the work of Pham *et al*., (2014) for finding experts reviewers and Yang *et al*., (2002) for providing recommendations in the form of actions that will help "naive" reviewers to switch into "expert" ones. Additionally, we study user evolution to find how reviewers improve over time and make action recommendations based on their current expertise level. As the first step, it is important to distinguish reviewers correctly that is reflective on the ability of reviewers to write helpful reviews.

To achieve this goal, we learn about user modeling and its applications covered in Section 2.2.

2.2 User modeling and its applications

User modeling is used to create user models in which observable information of a user is mined to infer unobservable information about a user (Frias-Martinez *et al*., 2006). User models can contain diverse information about a user or user groups such as user's domain knowledge, user's goals and plans, user's belief about the domain, specific preferences or interests, and user's attributes (Paris, 2015). For example, in a system that is acting as a librarian, user attributes could be "feminist" or "religious".

There are two distinct approaches of creating user models: (1) the user-guided approach and (2) the automatic approach (Fink *et al*., 1998). In the user-guided approach, also referred as the *explicit* approach, models are directly created using the information provided by the users themselves whereas in the automatic approach, the model creation process is controlled by a system which is unknown to user. Usually a user-guided approach is not preferred as users are unlikely to invest time to provide personal information unless it is compulsory, even though it is more direct and could obtain targeted response to specific questions. Furthermore, users may provide incorrect or inaccurate information regarding their interests or skills. In the automatic approach, also referred as the *implicit* approach, user information is derived from naturally occurring interactions between system and the user that user would have performed any way, without investing any additional time (Jameson, 2009). For example, the navigation pattern of users could help infer the behavior and interests of various subgroups of users based on the pages or categories they visited during their interaction with a website

(Dalamagas *et al*., 2007). User modeling is applied in diverse systems in order to model users to understand them better and eventually use this information to better serve them.

An example application of user modeling is the automation of *spoken dialogue systems* to individual users such as offering details about train arrival or flight departures via phone (Jameson, 2009). The users of these systems could be novice or experienced based on how experienced they are with the system. The dialogue systems should recognize these users and serve them accordingly: extensive and thorough explanation for novice users, and simple and quick sessions for experienced users (Jameson, 2009). Thus, a major goal of these systems is to perform user modeling to identify users' experience level in order to assist them better. Identification of user experience level is a difficult task and various systems use different measures ranging from simple to complex, to perform user experience identification. In spoken dialogue system, user models are represented based on the level of difficulty user faces on speech recognition when proceeding with specific dialogue (Litman and Pan, 2002). Many systems deploy a simple measure of identifying a new user as a novice whereas an old user who has interacted with the system in the past as experienced. This approach may be a good solution to new user cold-start problem. However, for amazon reviewers this may not work as McAuley (2013) points that "some users may already be experienced at the time of their first review". Therefore, reviewers should be observed for a certain period of time before labeling them as novice or expert in a review recommendation system.

A similar example is Kyoto City Bus Information System which deploys three measuring criteria for user model creation based on their dialogues with the system: (1) user's skill level in terms of using the system, (2) user's knowledge level in terms of

domain expertise, and (3) user's desire to complete the conversation quickly referred as urgency level (Komatani *et al*., 2005). Pieces of evidence such as the amount of information specified in each utterance by user; user's knowledge of exact bus stop location names; user's frequency of interruption before the system completed an utterance, etc. are collected implicitly to measure user's skill level. Similar to Kyoto City Bus Information System, in review recommendation system we can measure reviewer's expertise level by the amount information they post, product-specific keywords they use, review length, etc.

Another application is Web-based Intelligent Tutoring Systems (ITS) that performs user modeling to understand the knowledge and learning abilities of students by observing their interaction with the system (Suraweera *et al*., 2004). Student modeling or learner modeling is performed by observing various aspects of user characteristics: (1) user's domain knowledge based on the past and current interactions between the user and the system, (2) user's ability or motivation to learn, and (3) user's approach or the way of dealing a problem in hand (Jameson, 2009). For example in SQL-Tutor, based on the answers provided by student, the system provides feedback and helps the student determine a problem the student should attempt next. The student progress is measured implicitly based on his or her answers to perform student modeling (Jameson, 2009). Similar approach may be used in review recommendation framework to understand the progress a reviewer is making over time. However, reviewers are likely to consume and review diverse products ranging from science fiction novels to garment products which makes the progress harder to evaluate. Therefore this approach maybe suitable to

measure the progress of reviewers who tend to review similar products for example, super hero movie fans, gamers, book lovers, etc.

User modeling can be used to create reviewer models with different level of reviewing expertise. Paris (2015) defines a naïve user as one "who doesn't know about specific objects in the knowledge base and doesn't understand the underlying basic concepts" whereas an expert user as one who has domain knowledge and can relate to new objects based on his/her domain knowledge. Experts may not necessarily know about all the objects but has enough domain knowledge to either infer from a similar known object to understand a new object or to ask questions about it. Also a user does not have to belong to either an expert or a naïve class rather the user may belong somewhere in-between. "The level of expertise can be seen as a continuum from naïve to expert" (Paris, 2015). We extend the work of Paris (2015) in defining reviewers as experts or novices or any other classes. In Amazon.com, some reviewers have years of experience in purchasing and reviewing a specific product type. They know the exact information regarding product features or the depth of usage experiences they should share to effectively review the product. While other reviewers are relatively new and they may have no idea of the kind of information that they should share in their reviews to help other consumers.

2.3 User modeling process

The process of user modeling is divided into four steps (Frias-Martinez *et al*., 2006):

- Data collection
- Data preprocessing
- Pattern recognition
- Validation and interpretation

In this thesis, we follow all four steps of user modeling. Data collection is the process of gathering the information required for building user models. For our research, we used Amazon product reviews that is publicly available for research purpose. This Amazon product review data was collected by performing breadth-first-search on user-product-review graph until termination by McAuley *et al*. (2015). The details regarding number of users, reviews and products are presented in Table 4.1.

Data preprocessing involves getting rid of noise and inconsistencies present in the data. Data preprocessing also involves checking for impossible or unlikely values and missing values (Maglogiannis, 2007). In this phase, information regarding user identification and the user interaction with the system are extracted (Frias-Martinez *et al*., 2006). In our research we process Amazon product review data to identify various kind of data inconsistencies, if any, such as abrupt change in user count or review count. For example, the number of reviewers and product increased exponentially from 1997 to 2003 after which the growth has been more gradual. That 1997-2003 exponential growth doesn't reflect current growth rate, therefore we chose to ignore the unusual growth rate and cleaned the data accordingly. After removing inconsistencies, the data is transformed and aggregated with regard to reviewer information such as reviewer identity, total reviews written by the reviewer, average overall (rating) provided by the reviewer, active life (in months) of the reviewer and so on. Details on reviewer information is explained in Section 4.3.1. After data preprocessed, we use machine learning approach to recognize the patterns in user data which is covered in Section 2.2.1.

In the validation and interpretation phase, patterns discovered from pattern recognition phase are analyzed and interpreted based on the feature values of each

patterns (Frias-Martinez *et al*., 2006). The interpretation involves use of domain knowledge and visualization. Validation tests the usability of the knowledge obtained. For Amazon.com reviewers, interpretation involves labeling the classes as experts or novice based on their feature values. Validation involves training and testing using supervised approach.

2.3.1 Pattern recognition using unsupervised approach

Pattern recognition is the process in which computer program discovers patterns of the objects it has seen before, for example chronological or spatial pattern (Anzai, 2012). In other words, “pattern recognition is a process of generalizing and transforming representations” (Anzai, 2012). Pattern recognition is performed by various machine learning techniques which may be either supervised or unsupervised depending upon the dataset. Clustering algorithms such as *K*-means clustering, *X*-means clustering, and correlation clustering are major examples of unsupervised algorithms whereas classifiers such as decision trees, Naïve Bayes classifier, and neural networks are examples of supervised algorithms. The choice of using supervised or unsupervised approach depends on whether the instances in dataset are labeled or not (Maglogiannis, 2007). If all the instances have known label then supervised approach is used and if the instances are unlabeled then unsupervised approach is used (Maglogiannis, 2007). For our research, the reviewer data has no label and we use unsupervised learning to discover the unknown patterns. Amazon.com reviews data contain Average Helpfulness and Average Overall which may be used as labels. They are continuous "label", not nominal. Some might suggest discretizing the continuous labels to represent different levels of expertise which may be interpreted as "label". For example in the case of Average Overall, ratings like:

less than 1 star, between 1 and 2 stars, between 2 and 3 stars, between 3 and 4 stars, and between 4 and 5 stars as a way to represent different levels of expertise. But we argue that expertise does not have to be a 1-to-1 mapping with Average Overall, or with Average Helpfulness rather expertise is a function of multiple features including Average Helpfulness, Average Overall, Active Month, Review Frequency and so on. Hence we consider the reviewer data as unlabeled and have to resort to unsupervised learning to discover reviewer expertise clusters.

A clustering algorithm, which is an unsupervised learning procedure, groups a set of objects, i.e., items or users in such a way that similar objects are grouped within a same group and are dissimilar to the objects in another group (Gan *et al*., 2007). Similarity coefficients are used for quantitatively describing the similarity between two clusters. For numerical data similarities between two objects are measured using distance metrics like *Minkowski distance*, *Mahalanobis distance*, and *average distance* or combination of these distances (Gan *et al*., 2007). Minkowski distances are the standard metrics for geometrical problems (Strehl *et al*., 2000). The advantages of using Minkowski distances are that they are easy to compute and allow scalable solutions to clustering problem (Gan *et al*., 2007).

For two objects, $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, *Minkowski distance* is defined as,

$$d(X,Y) = \sqrt[q]{\sum_{i=1}^{n} |x_i - y_i|^q},$$

where n is the dimension of the object and x_i, y_i are the values of ith dimension of the object X and Y respectively, and q is a positive integer. When $q = 1$, d is *Manhattan*

distance and when $q = 2$, d is *Euclidean distance*. Among others Euclidean distance may be the most common distance that has been used for numerical data (Gan *et al*., 2007).

Cluster analysis has been used widely to understand the target users in diverse fields. Clatworthy *et al*., (2005) reviews number of cases in health psychology where cluster analysis is used to address various theoretical and practical problems. Clustering is mainly used to identify people or groups at risk of certain medical condition in order to assist them with the required medical service (Gan *et al*., 2007). Clustering has also been used in market segmentation research (Wedel *et al*., 2012). In market segmentation research, clustering is used to assign potential customers to homogeneous groups based on various characteristics such as cultural, geographic, demographic, and socio economic factor. For Amazon.com reviewers, pattern recognition involves clustering of similar reviewers into different classes with interpretable differences. By similar reviewers, we mean reviewers who show similarities in their characteristics such as reviewing frequency, review text length, review rating (overall), active period, and review helpfulness.

As stated earlier *K*-means clustering, *X*-means clustering, and correlation clustering are some of the major clustering algorithms employed for pattern recognition for unlabeled data. *K*-means clustering is the most popular and simplest hierarchical clustering algorithm since it was proposed 50 year ago (Jain, 2010). Its popularity is largely due to the ease of implementation and empirical success. However there are three major shortcomings of *K*-means algorithm such as (1) poor computational scalability, (2) the number of clusters denoted by *K* has to be supplied as a parameter, and (3) local

minimum convergence (Pelleg *et al*., 2000). *X*-means clustering algorithm overcomes the first two shortcomings which makes *X*-means algorithm the best fit for our dataset.

There are some open questions about our research goal that should be understood in order to choose suitable clustering algorithm for our dataset which contains all kinds of reviewers ranging from new to experienced and novice to experts. Two important considerations are:

- Is there any fixed number of clusters that reviewers should identify to? No. Reviewers can be grouped into any number of clusters depending upon the inherent patterns in the reviewer features. The number of clusters may depend on how diverse reviewers are. We don't want to impose any restrictions on how many clusters of reviewers should be created. The clusters will be interpreted based on their expertise level ranging from novice to expert. The level of expertise is seen as a continuum from novice to expert (Paris, 2015), thus we cannot predetermine the number of clusters.
- Are computationally slow learning acceptable? The acceptability of slow learning algorithm is questionable for two main reasons: (1) large number of online reviews and (2) changeable nature of reviewer clusters.
    - Ever increasing number of reviews and reviewers in Amazon.com makes it more important to employ computationally faster learning algorithm.
    - Given that clusters are likely to change over time and that new clusters will likely be needed after new reviews are posted or new reviewers joins, the need to employ faster learning becomes crucial.

The *X*-means algorithm doesn't require us to provide the number of clusters and is computationally efficient which makes it suitable for large datasets (Pelleg *et al*., 2000). Therefore we choose *X*-means clustering algorithm to perform pattern recognition in Amazon.com reviewers. In the end of this phase, a structural description of what the system learned about user behavior and user interest is obtained as output (Frias-Martinez *et al*., 2006).

Note that soft clustering such as fuzzy clustering assigns each data into multiple clusters instead of one single cluster (Dunn, 1973). This feature may be particularly important in our case because we may want to know the degree or percentage of belonging of a reviewer to each of the clusters and provide recommendations to the reviewer accordingly. However, *X*-means also provides confidence interval of each reviewer which is a measure of how strongly the reviewer belongs to the cluster. Also, similar to *K*-means, fuzzy clustering should be provided with the number of clusters as a parameter. Hence ultimately we choose *X*-means over fuzzy clustering as *X*-means doesn't require us to provide the number of clusters for a given dataset.

2.3.2 Validation and interpretation using supervised approach

Interpretation is the process of analyzing and interpreting the structures discovered from pattern recognition phase (Frias-Martinez *et al*., 2006). For our research, after performing cluster analysis covered in Section 2.2.1, all the instances belong to a certain cluster or group. We label these clusters based on their behavior for example a cluster of reviewers whose reviews have been highly helpful to other customers are labeled as 'experts' whereas a cluster whose reviews have not been helpful to other customers are labeled as 'novices'.

Validation is testing the credibility of the structures discovered from pattern recognition phase (Frias-Martinez *et al*., 2006) and it is usually carried out by (1) creating a predictive model and (2) performing model validation of the prediction capability of the model. Supervised approach is used to build a predictive model of the class labels based on the predictor features (Kotsiantis *et al*., 2007). The predictive model is a classification function that maps input or instances to class labels.

Classification is the problem of predicting or automatically labeling the class of a new instance on the basis of training data, which contains the instances whose classes or labels are known. There are many example applications of such classifiers. Beck *et al*., (2003) used a classifier to predict if a student would request for help in an intelligent tutor for reading. The training data set contained students description based on their interactions with the tutor. The classifier was able to predict if a student would click on a particular word for help with 83.2% accuracy. Kwapisz *et al*., (2013) created a predictive model that recognizes the activity a person is engaged in based on the cell phone accelerometers. Smart cell phones have acceleration sensors i.e., accelerometers which can be used to perform activity recognition such as walking, jogging, sitting, climbing stairs and so on. This information is used as a training data to create a predictive model for activity recognition. In our research, a classifier is trained using a dataset of reviewers who have been classified into different classes based on their expertise level to write helpful reviews. After performing cluster analysis, all the instances are labeled and belong to a certain class. This data is used as a training set to train a classifier that maps unlabeled reviewers to a class. The trained classifier can then be used to predict the class of a new reviewer based on their expertise level.

There are mainly three different types of classification algorithms such as decision trees, neural networks, and ensemble learning. The decision tree approach is one of the most widely used approaches to represent classifiers. It originated from the field of decision theory and statistics; however, it is very popular in other fields such as data mining, machine learning, and pattern recognition (Deepti *et al*., 2010). Decision tree is a classifier in the form of a tree structure where each node is either (1) a leaf node or (2) a decision node. A leaf node represents the decision outcomes (class) whereas a decision node represents a test to be performed on one or more attributes. A decision node may have two or more branches corresponding to a range of values. These ranges of values must give a partition of the set of values of the given characteristics. The decision trees have many advantages. Decision trees (1) are easy to understand, (2) can be easily converted to a set of production rules, and (3) can classify both categorical and numerical data, and (4) do not have *a priori* assumptions about the nature of the data (Zhao and Zhang, 2008). However, decision tress have some disadvantages. For instance, they are unstable which means that slight variations in training data can result in different attribute selection at each point within the tree. This can make a significant change as attribute choices affect all the descendent subtrees. Although decision trees have some disadvantages, they are suitable for our research for three important reasons: (1) time efficiency, (2) easy to understand, and (3) easy conversion into production rules. Unlike neural networks, decision trees can be reduced to set of rules which is important in our case because we want to find which features are used to perform partitioning at different levels. This information will help us see which feature has more predictive weightage

over another. In other words, this approach better helps to understand the feature that is used for classification of reviewers at different level.

2.4 Sentiment Analysis

Sentiment analysis, also referred as opinion mining is defined as "the analysis of people's opinions, sentiments, evaluations, appraisals, attitudes and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes" (Liu, 2012). With a growing availability of user-generated text in social media, blogs, and online reviews, there are new opportunities to seek and understand users opinions (Bo & Lee, 2008).

There are many examples of sentiment analysis applications in diverse fields from predicting stock trading to election results. Zhang and Skiena, (2010) researched the effects of company-related news published in quantitative media like blogs, news, etc. on their stock trading. They studied sentiment-oriented equity trading based completely on blog/news data. Tumasjan *et al*., (2010) used sentiment analysis to predict election results from Twitter in the context of German federal election. Political sentiment collected from tweets that mentioned a political party reflected the offline political landscape. McGlohon *et al*., (2010) used product reviews to rank products or merchants. Reviews tend to contain reviewer's individual biases and the reviewer is likely to carry the same "bias" around the products they rate. Sentiment analysis helps to understand these biases then measure the true quality of product or merchant. Hai *et al*., (2011) used sentiment analysis to understand opinion features on online reviews. Opinion words represent explicit features which are used to identify implicit features in a sentence. Explicit and implicit opinion features helps to produce finer-grained understanding of online reviews.

Sentiment analysis of online reviews in Amazon.com can help to understand reviewers opinion on the consumed product. We extend the work of McGlohon *et al*., (2010) and Hai *et al*., (2011) of analyzing online reviews. *We try to understand the relation between the tone of online reviews and their perceived helpfulness.* For example, reviews with positive opinion may be more effective for certain products whereas negative opinion may be effective for other products. This will help to understand how customers perceive positive, negative or moderate opinions for different products.

2.4.1 VADER

VADER stands for Valence Aware Dictionary for sEntiment Reasoning. It is a simple rule-based model for general sentiment analysis tool which was created from a generalized, valence-based, human-curated gold standard sentiment lexicon that is especially attuned to microblog-like contexts (Ribeiro *et al*., 2015). VADER combines lexical features with five generalized rules to incorporate grammatical and syntactic features that humans use to express sentiment intensity such as (1) punctuation namely exclamation point, (2) capitalization like use of ALL-CAPS, (3) degree modifiers like degree adverbs for example *extremely*, *marginally*, and so on, (4) use of contrastive conjunction such as *but*, and (4) examining tri-gram preceding a sentiment-laden lexical feature for example "*The service here isn't really all that great*" (Hutto & Gilbert, 2014). The input to VADER should be in the form of texts such as tweets, reviews, etc. The output received from VADER is the measurement of sentiment *polarity* such as positive, negative, and neutral; and sentiment *intensity* on the scale of -4 to +4. The *intensity* is measured as -1 to -4 for slightly, moderately, very and extremely negative, 0 for neutral, and 1 to 4 for slightly, moderately, very and extremely positive.

VADER has been proven to perform as well as (and in most cases, better than) than eleven other highly regarded state-of-practice tools such as LIWC, ANEW, the General Inquirer, SentiWordNet, and machine learning oriented techniques relying on Naïve Bayes, Maximum Entropy, and Support Vector Machine algorithms (Hutto & Gilbert, 2014). VADER outperforms aforementioned sentiment analysis tools in the analysis of social media texts from Tweets, Amazon product reviews, and NY Times Editorials (Hutto & Gilbert, 2014). Sentiment analysis using supervised machine learning models tend to be more accurate as they are trained using the same corpus which they later classify. Socher et al. (2013) talks about a sentiment tree bank, a recursive deep model that is reported to outperform the state-of-art supervised machine learning model. However, the results of VADER are on par with sentiment tree bank (Hutto & Gilbert, 2014). Further, VADER is quick and computationally economical. The lexicon and rules used by VADER are directly accessible and can be modified as needed. Among many advantages of using VADER, its accuracy of analyzing Amazon product reviews (Hutto & Gilbert, 2014) is the main reason we choose to use VADER for performing sentiment analysis in our research.

Chapter 3
Methodology

The aim of this thesis, as stated in Chapter 1, is to differentiate reviewers based on the quality of their reviews to identify different classes of reviewers such as expert, novice, etc. and to devise an approach to leverage the experts' behaviors to help train novice reviewers effectively and efficiently. In this chapter we present a higher level overview of the online review recommendation system framework which utilizes clustering and decision tree based classifier to differentiate various classes of reviewers. We cover the methodology in two main parts: (1) differentiating reviewers based on review quality and (2) recommendation system framework. We first try to understand reviewers by analyzing the reviews they have posted so far. Once we have the general idea about reviewers behavior and their reviewing skill, we then develop a framework to recommend actions to novice reviewers by observing the closest expert behavior.

3.1 Modeling different categories of reviewers based on review quality

Reviewers have different reviewing skills based on their expertise. To differentiate reviewers expertise level, in other words, to measure reviewers' skill to write quality review, we start by analyzing their reviews. First, we define review quality and find factors that affect review quality. Second, we extract feature set that can help us understand reviewer behavior. Third, we model the reviewers based on their features to determine their expertise level. We will discuss aforementioned three steps in following sections.

3.1.1 Defining review quality

As stated in Chapter 1, definition of review quality is subjective and depends largely on the perception. Review quality has different definitions based on customer and retailer perspectives. The feature set that defines review quality may be different from buyer and seller perspectives. From a retailer perspective, review quality is a measure of their product sales, that is, a review which helps to increase product sales is considered as good quality review (Chevalier and Mayzlin, 2006; Lee *et al*., 2008). However, for our research, we focus on customer perspective to define review quality and therefore focus on review features that reflect customers perception. As stated in Chapter 1, customers consider a review as good quality if it helps them to make an informed decision which may or may not lead to the purchasing of product.

Online retailers including Amazon have been using "helpfulness" as the primary way of measuring consumers' evaluation of a review (Mudambi and Schuff, 2010). In the end of each review, Amazon.com asks if the review was helpful to the reader. Helpfulness of a review is a number of up-vote the review receives from customers who like the review. Helpfulness has been interpreted as a measure of customers' perceived value in decision-making process (Mudambi and Schuff, 2010; Otterbacher, 2009). Therefore ***helpfulness*** is an important review feature that we use to measure customers perception of review quality.

3.1.2 Feature set

There are different types of information available in Amazon product reviews which may be regarding product, reviewer, and review. When extracting a feature set, it is important

to note that the features should somehow help us to assess review quality which in turn will assess reviewers' expertise level.–Below are example features under each category:

Product information: Information of a product such as product categories like electronics, books, etc. As discussed previously in Chapter 2, Zhang *et al*. (2010) points that product reviews are evaluated differently based on consumers' consumption goals. For some products, consumers want to identify useful information for achieving outcomes, referred as *promotion consumption goal*, whereas for other products they want to identify useful information for avoiding undesirable outcomes, referred as *prevention consumption goal* (Zhang *et al*., 2010). The perceived helpfulness of a review may be different for different product types. Hence for our research, we include diverse product categories which may fall into any of the abovementioned product types.

Reviewer information: Information related to a reviewer such as reviews count of the reviewer, reviewer active life span, etc. Reviews count of a reviewer is calculated from the number of reviews posted by the reviewer. Similarly, review posted timestamps are used to keep track of reviewer active life span in months. These information represent reviewer's personal behavior and measure the reviewing frequency of an individual reviewers. Intuitively, we think that these pieces of reviewer information might help to understand reviewers and more accurately assess their expertise level. For example, a reviewer who consistently and regularly contributes helpful reviews should be considered more expert than a reviewer who only occasionally contributes helpful reviews.

Review information: Information related to a review such as review depth, review extremity, etc. Review depth is the length of a review and review extremity is the rating or overall assigned by reviewer to the product being reviewed. Mudambi and Schuff

(2010) point that review extremity and review depth affect the perceived helpfulness of the review. Since perceived helpfulness is the measure of review quality, these features will affect review quality and in turn affect reviewers' expertise level.

3.1.3 Reviewer modeling

Reviewer modeling is performed to understand the reviewing behaviors of reviewers and identify common patterns in their behavior. Reviewer behavior is represented by different features listed in Section 3.1.2 which may be related to product, review, and reviewer. In order to understand reviewer behavior, we use the following different features of a reviewer: (1) total number of reviews posted by them, (2) total time in months they have been reviewing, (3) average rating (overall) the reviewer assigns to the product being reviewed, (4) average length of reviews posted by the reviewer, and (5) average helpfulness received by their reviews so far. Based on the similarity of aforementioned features, we find common patterns in reviewers behavior by grouping similar reviewers together.

We analyze each group of reviewers, to understand them better. The average helpfulness of each group is the quality of review the group writes. Based on the review quality and other features we label each group appropriately. We use average helpfulness as a primary measure to differentiate between expert reviewers and novice reviewers. However, past research shows that features like product type, review length and review extremity also have an impact on helpfulness (Mudambi and Schuff, 2010). Therefore, in order to differentiate reviewers, we look beyond helpfulness by analyzing average review length, average review extremity and product type as well.

3.1.4. Overview

The goal of modeling different categories of reviewers based on review quality is accomplished using different techniques. First, for feature sets extraction, some features are readily available such as review post timestamps whereas others are derived from existing features such as average review length written by a reviewer is calculated by averaging the review text length of all the reviews posted by them. Details of feature extraction process is covered in Section 4.3.1. Second, clustering is used to group the reviewers into different classes. Clustering of reviewers is followed by labeling each reviewer class based on the helpfulness of their reviews and other feature set. Details of the clustering process and reviewer class labeling are covered in Section 4.3.2. Third, decision trees are used to classify a new reviewer into an appropriate class and validate the classification process. Decision trees help us to understand how classification is performed and which features are used to perform classification at different levels. Classification process using decision tree is covered in Section 4.4.

3.2 Recommendation system framework

In this section we present an architecture that recommends actions to novice reviewers by learning from expert reviewers. The recommendation system trains novice to follow the action sequence of expert in order to improve their reviewing skill. To achieve this, we first understand if reviewers change over time with respect to their reviewing skill. If they change over time, we want to know how they change. This will provide insights on how reviewers are evolving on their own and in which phase of their evolution should our recommendation system framework facilitate. Second, we perform sentiment analysis on the review text to understand the tone used by different classes of reviewers. We want to

understand if different classes of reviewers use a different tone and how that affects on review helpfulness. Third, we present an architecture that recommends actions from closest expert to train the reviewer who is lagging behind.

3.2.1 Reviewers evolution

McAuley and Leskovec (2013) point out that users evolve over time in terms of taste and properties of products. This process of user evolution or change in users' tastes takes place with knowledge, maturity and experience and is referred as *personal development* of users. McAuley and Leskovec (2013) prove that users or reviewers with similar level of experience will rate products in similar ways, even if their rating are temporally apart. Rating of a product is reflected in review text. For example, the review text of a highly rated product will have more positive descriptions of the product while the review text of a lowly rated product will have more negative descriptions of the product. Since reviewers' experiences affect their ratings, in this thesis, we try to find if experience affects their reviewing skills as well.

For our research, if reviewers evolve overtime, we want to look into the trend of their evolution with regards to their reviewing skill or their ability to write quality reviews. As stated in Section 3.1.1, we measure review quality in terms of the perceived *helpfulness* received by the review. For a reviewer, their reviewing ability is measured in average *helpfulness* received by their reviews. The evolution, if any, may be directed either upward when reviewers start to post more helpful reviews over time or downward when reviewers start to post less helpful reviews over time. If there is no evolution, the trend will be more or less constant, or, in other words, average helpfulness of reviews posted by reviewers will be constant over time.

In this thesis, we focus on finding if reviewer classes evolve to form a trend that follows a pathway. With evolution, novice reviewers may evolve into expert reviewers or vice versa. McAuley and Leskovec (2013) have done a similar research on users experience level and their findings indicate that there are three types of users: (1) users who evolve into experts after progressing through all levels of experience, (2) users who never become experts, and (3) users who start as a expert from the very beginning. We follow a similar approach of treating experience as a function of time, to find how different classes of reviewers evolve from one class to another. These findings can help us to understand the evolution patterns of each reviewer class and treat them accordingly when developing our recommendation system framework.

3.2.2 Review sentiment analysis

Reviewers share their experience with the product via review text. In other words, a review text expresses its author's (i.e., reviewer's) opinion. There is a linguistic variability in review texts as they express different opinions, questions, and products (McAuley and Yang, 2016). Sentiment analysis helps to understand the opinion polarity in a review text. This will provide insights into whether different reviewer classes—as proposed in earlier sections—have different types of sentiment polarity.

Note that there are different ways to express same opinion, some of which may be pleasing to the readers whereas others may not. As stated in Chapter 2, depending on the product consumption goal, consumers are inclined towards positive reviews for some products whereas negative reviews for other products (Zhang *et al*., 2010). Therefore, perceived helpfulness of a review may be directly related to linguistic difference in review text. We have to consider the effect of language used by different classes of

reviewers—novice and expert. There is a strong relation with "expertise" from the light of linguistic development (Romaine, 1984). Experts may have a more pleasing writing style than novices that makes their reviews perceived as more helpful. A pleasing writing style may indicate providing positive/negative information, using more/less pronouns, writing personal reviews by using "I" instead of "We", and so on. To explore and answer these questions, we thus perform sentiment analysis in review text.

3.2.3 Review recommendation system

In this Thesis we propose a recommendation system framework that provides recommendations in the form of reviews to help reviewers to improve their reviewing skills and write better quality reviews. Traditionally, *recommendation systems* provide recommendations on web pages, videos, movies, music, and books (McDonald *et al*., 2000). Recommendation systems have been helping customers make decisions on which products to purchase on E-commerce websites (Schafer *et al*., 2001). Note that our review recommendation system is different from traditional recommendation systems in two major ways:

1. Traditional recommendation systems provide suggestions for items to be of use to a user (Shapira *et al*., 2011). The suggestions are aimed at supporting the users in various decision-making processes, such as which item to buy, which music to listen, or what news to watch. Unlike traditional recommendation system, our recommendation system framework generates action recommendations such as which review to post or what kind of review to post for reviewers. The recommendations are not in the form of items or products but rather ***in the form of actions that the reviewer should perform***. The recommendations are actual

review's text that could potentially improve the reviewing skill of the reviewer who reads the text. More formally the review recommendation problem can be stated as: Let *R* be the set of all reviewers and let *V* be the set of all possible reviews that can be recommended. The space of *V* can be very large ranging to hundreds of millions of reviews and space of *R* can range in the millions as well in some cases. For each reviewer $r \in R$, we want to choose a review $v' \in V$ to help *r* improve the quality of his or her future reviews.

2. In a traditional recommender system for item recommendations (such as movie, music, etc.), the usefulness of an item to a user is usually represented by a *rating* which is a measure of how much the item is favored by the user (Adomavicius *et al*., 2005). However, in our case, the usefulness of a recommended review to a reviewer is measured by how *helpful* the subsequent or resultant review written and posted by the reviewer after having read the recommended review. Such a helpfulness measure is, in turn, computed from customers' perspective, as it is derived from how they receive or rate the posted review. More specifically, for a reviewer r, who uses a (recommended) review v_1 to posts a review v_1', the success of v_1 is determined by the number of up-votes or *helpfulness* generated by v_1', since we define the review quality in terms of *helpfulness* (Section 3.1.1).

Chapter 4
Understanding Reviewers

As stated in Chapter 1, one of our main goals in this Thesis is to differentiate reviewers based on the *quality* of their reviews to identify different classes of reviewers. In this Chapter, we try to understand reviewers by first differentiating and then predicting reviewers based on their review quality via user modeling. In building a model that is capable of differentiating and predicting a reviewer class, we have to perform following processes:

1. Data collection mechanism (discussed in Section 4.1 in Chapter 4)
2. Data preprocessing mechanism (discussed in Section 4.2 in Chapter 4)
3. Data clustering (discussed in Section 4.3 in Chapter 4)
4. Data classification (discussed in Section 4.4 in Chapter 4)

We discuss each of the implementations in following sections. While we perform the aforementioned experiments, we pursue the following series of objectives:

1. **Objective 1**: Demonstrate that reviewers can be either expert or novice by performing data clustering and then doing data analysis to identify attributes that make them expert or novice. We will use the quality and quantity of reviews as metrics to define expert and novice reviewers. Differentiating different classes of reviewers will help to further understand the behavior of each class over time and over different product type.

2. **Objective 2**: Demonstrate that a number of features like review length, overall (rating), helpfulness, etc. affect review classification by developing decision tree for data classification to find features that differentiates clusters from one another. Understanding which feature plays more important role than other to perform classification will help us find the features that are more important than other.
3. **Objective 3**: Demonstrate that reviews are valued differently across different product categories by performing linear analysis on reviews of diverse product categories such as Books, Electronics, Cellphones and accessories, Health and personal care, Grocery and gourmet foods, Office products, and Baby. Understanding that reviews are valued differently across different product types will help us identify salient features for each category.

4.1 Data collection mechanism

The data for our research was extracted from *Amazon.com* web store. Amazon website stores information from its online interactions with buyers. This information includes the details of each review written by buyers also referred as reviewers in this context. It contains (1) the review text, (2) a reviewer id which is in alphanumeric format, (3) the review post timestamp, (4) the product id in alphanumeric format, (5) *overall* which is the rating that the reviewer assigned to the product he/she purchased, and (6) helpfulness which is the number of votes that the review received from other users who found the review helpful.

This Amazon product review data was collected by performing breadth-first-search on user-product-review graph until termination by McAuley, *et al*. (2015). It is available online in one-review-per-line in loose JSON format for academic research

purpose. The dataset contains product reviews from Amazon, including 142.8 million reviews spanning from May 1996 to July 2014. The reviews are separated into 24 different product categories such as Automotive, Beauty, Books, Digital Music, Electronics, and so on.

To summarize what was previously discussed in Chapter 2 and Chapter 3, reviewers seek and write different information based on product type. Zhang *et al*. (2010) points that product reviews are evaluated differently based on consumers' consumption goals. For products associated with *promotion consumption goal*, consumers want to identify useful information for achieving outcomes, whereas for products associated with *prevention consumption goal*, they want to identify useful information for avoiding undesirable outcomes (Zhang *et al*., 2010). The perceived helpfulness of a review may be different for different product types. Hence for our research, we include diverse product categories which may fall into any of the abovementioned product types. To generalize reviewers differentiation approach, we choose to examine reviewers' characteristics beyond single product category. Doing so demands a large amount of training data in diverse categories, which will strengthen the novelty of user modeling approach we propose. We choose nine different categories for our research whose characteristics such as number of users who have provided at least one review; number of products; number of reviews; and consumption goal of each category are shown in Table 4.1.

Category	Users	Product	Reviews	Goal
Books	8,201,127	1,606,219	25,875,237	Promotion Consumption
Electronics	4,248,431	305,029	11,355,142	Promotion Consumption
Cell Phones and accessories	2,296,534	223,680	5,929,668	Promotion Consumption
Grocery and gourmet food	774,095	120,774	1,997,599	Prevention Consumption
Health and personal care	1,851,132	252,331	2,982,326	Prevention Consumption
Office Product	919,512	94,820	1,514,235	Promotion Consumption
Baby	19,445	7,050	160,792	Prevention Consumption
Beauty	1,210,271	249,274	2,023,070	Prevention Consumption
Pet supplies	740,985	103,288	1,235,316	Prevention Consumption

Table 4.1: Dataset statistics for our experiment

Review data of all the above product categories are examined individually. Individualizing experiments for each product categories allowed us to model expert and novice differently based on the product category, which is one of the crucial parts of our user modeling.

4.2 Data preprocessing mechanism

The Amazon product review data obtained online is parsed using python code and is then run through various test to get the better understanding on the data. We looked at user distribution, product distribution, and review distribution, all of which will be discussed in detail in Section 4.2.1. We observed that the review data was imbalanced and at the same time contained large proportion of inactive users, which will be discussed in Section 4.2.1 and Section 4.2.2 along with the solutions, we came up with to deal with aforementioned problems.

4.2.1 Data cleaning to address existing data imbalance

To get a closer understanding of Amazon product review data, we started by counting number of users referred as user count, number of products referred as product count and number of reviews referred as review count in chronological order for every month. The distribution was imbalanced, growing exponentially for first few years and then growing linearly after that until 2013.

Amazon books review data contains reviews from 1996 to 2014. Below is a graph of review count; user count and product count in log with respect to time in month.

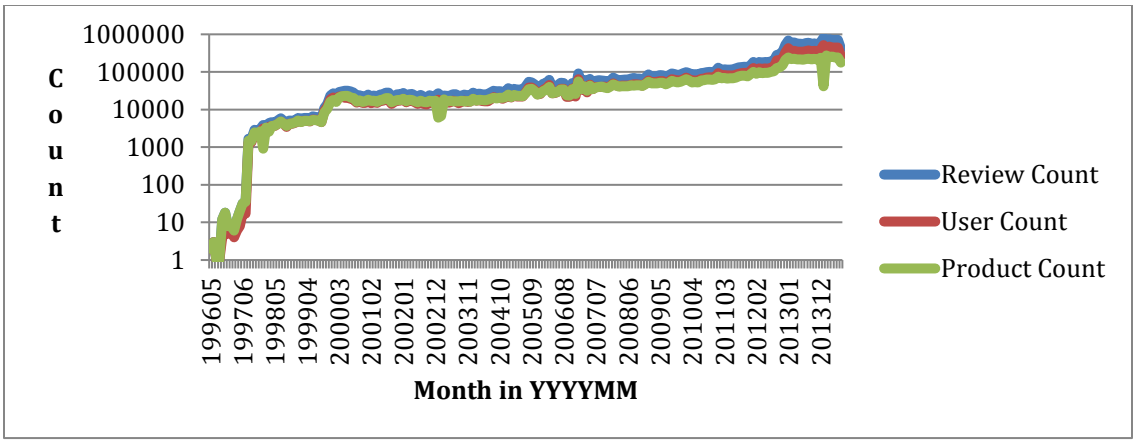

Figure 4.1: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Books" category.

In Figure 4.1, user, review and product count grow exponentially before year 2000, followed by more gradual and consistent growth from 2003 to 2013. One of the probable explanations for the exponential growth could be the Internet traffic growth, which was close to doubling between 1997 and 2002 (Odlyzko, 2003). Considering that this exponential growth occurred in distant past, more than a decade ago and is not an accurate representation of recent growth, we have decided to ignore the data before 2003. At a closer look we can observe that product count in January 2014 decreases abruptly by 82% whereas user and review count don't decrease along with product count, which is highly unnatural considering the high correlation, specifically ~0.99 between these three counts for last 10 years from 2003 to 2013. Also, user, review and product count has a decreasing trend in the last month that is July 2014, which is probably due to data incompleteness for the month. Considering the data inconsistencies due to data incompleteness in 2014, we choose to ignore the data of year 2014, and include only year-round complete data spanning from January 2003 to December 2013. Figure 4.2 shows review count; user count and product count in log with respect to time in month from January 2003 to December 2013.

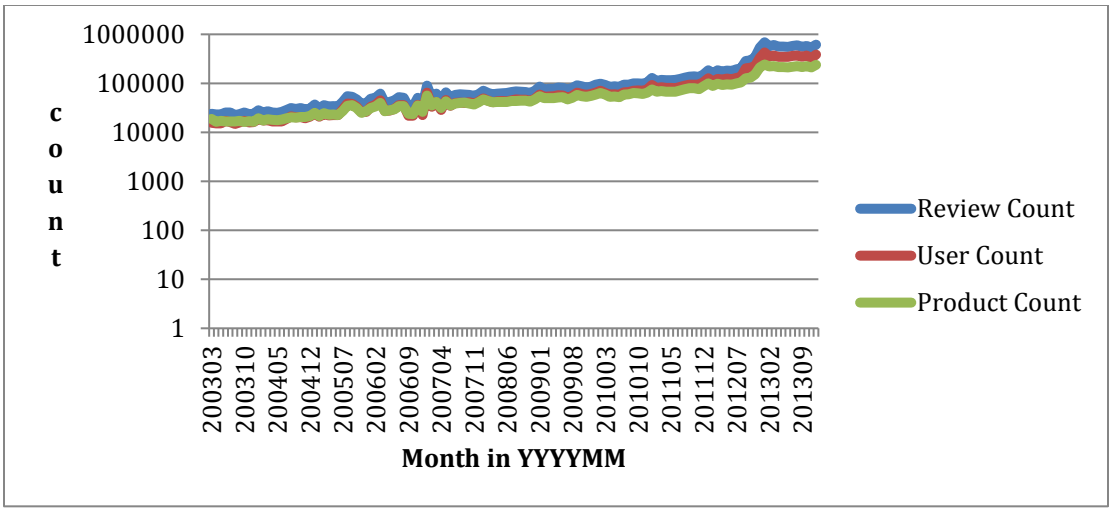

Figure 4.2: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) after cleaning the data in "Books" category.

After cleaning, Amazon books review data from Jan 2003 to Dec 2013 looks more gradual and is more correct representation of recent growth trends in user, review and product count.

We repeated this process of data observation, analysis followed by cleaning for Amazon product review data of the nine product categories listed in Table 4.1 individually. The result of this process is provided in Table 4.2, which displays the duration in year before and after the data was cleaned to address data imbalance.

Category	Year before cleaning	Year after cleaning
Books	1996 to 2014	2003 to 2013
Electronics	1998 to 2014	2003 to 2013
Cell Phones and accessories	1991 to 2014	2003 to 2013
Grocery and gourmet food	2000 to 2014	2003 to 2013
Health and personal care	1998 to 2014	2003 to 2013
Office product	1998 to 2014	2003 to 2013
Baby	1998 to 2014	2003 to 2013
Beauty	1998 to 2014	2003 to 2013
Pet supplies	1998 to 2014	2003 to 2013

Table 4.2: Data statistics of before and after data cleaning to address data imbalance

After cleaning Amazon product review data, the chronological distribution is more gradual making the data balanced in terms of user count; product count and review

count in chronological order for all nine product categories which can be observed in Appendix, Section A.

4.2.2 Data cleaning to remove large proportion of inactive users

Active users are the total number of reviewers who have been reviewing for a certain number of months, which may or may not be continuous. Active users, in this context have a review history for certain number of months, which can be used for better understanding the reviewers. There is a large number of users who are active for very few months that makes it harder to find their characteristics to understand their reviewing pattern. These users are called inactive users, as their review experience is low in terms of number of month they have reviewed. These inactive users may or may not be expert reviewers. We have decided to remove them from consideration.

In an ideal case, the number of active users increases gradually as the number of month increases, which means that the old users keep reviewing consistently as well as new users have started reviewing with time. However, from our findings we observe that there are many users who discontinue to review with time. These inactive users usually make a large proportion of the user pool and if used in user modeling tend to dilute results, as this large pool of inactive users has very less review history. As this data will be used as a training set for clustering and classification we have chosen to include active users with substantial review history to strengthen the validity of our clustering and classification.

To identify the threshold value on number of months that differentiates between active and inactive users, we observe the number of users active over time and how the user count changes.

For this, as detailed in Table 4.3, we used Amazon books review data to plot top 15 values of total number of users active for respective month count referred as user count and calculated the slope of user count for the respective month and higher. For example, slope when month count is 2 indicates the slope of user count corresponding to month count 2, 3, 4, and higher. The slope is an indicator of the rate at which user count is increasing or decreasing with respect to month count. From Table 4.3, we observe that there is a drastic change in the number of user count over month count for months 1 through 5. The month count corresponding to the point from where the change in slope is gradual is picked as the threshold that differentiates active users from inactive users.

Month Count	User Count	Slope
1	4509951	**17563**
2	924735	**4927.1**
3	347701	**2339.1**
4	168642	**1365.9**
5	94484	**893.2**
6	57812	627.7
7	38547	465.2
8	26397	355.9
9	19153	281
10	14222	226
11	10746	184.8
12	8195	153.8
13	6575	129.8
14	5182	110.3
15	4199	95.1

Table 4.3: Distribution of User count, with respect to their active month (i.e. number of month they have posted a review)

Looking at Table 4.3, user count vs. month count, we observe that there are 5951029 (~95%) users who are active for less than 5 month and 315322 (~5%) users who are active for 5 month or more. In this case, month count 5 is a threshold that differentiates active and inactive users i.e. 95% of the users are inactive, as they have reviewed for 4 month or less and 5% of the users who have reviewed for 5 month or more are active as they have consistent reviewing frequency.

To get a visual representation of the above finding we plot Figures 4.3 and 4.4 to see the graphs of user count with respect to month count before and after removing inactive users i.e. users active for less than 5 month.

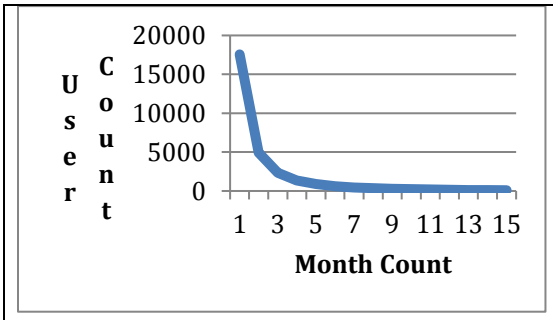

Figure 4.3: User count vs. Month count before removing users active for less than 5 month

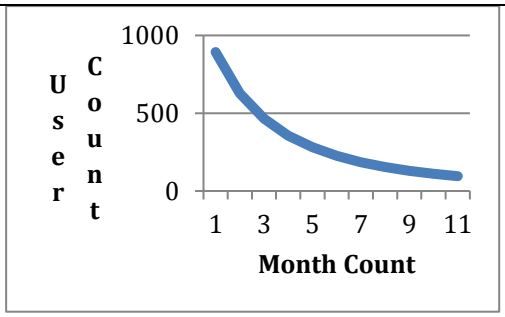

Figure 4.4: User count vs. Month count after removing users active for less than 5 month

From Figure 4.3, we can observe that the graph descends quickly till active month is 5 and descends gradually after that i.e. the number of users who are active for 5 month or more are more-or-less linear with respect to month count. We can also see that the number of users who are active for 4 months or less grow (or shrink) almost exponentially. From Figure 4.4, we can say that the user count decreases gradually with month count after removing inactive users that is users active for less than 5 month. This trend is balanced and is more correct representation of active user count with time.

We repeat this process to find the threshold month count in order to differentiate active and inactive users in each of the nine product categories listed in Table 4.1 individually. The details of month count; user count; and slope for all nine -product categories which can be observed in Appendix, Section B.

Table 4.4 details the threshold month count for all the nine categories. It can be observed that threshold month count is not uniform and is different for different product categories.

Product categories	Threshold month count

Books	3
Electronics	3
Cell Phones and accessories	4
Grocery and gourmet food	3
Health and personal care	3
Office product	3
Baby	4
Beauty	3
Pet supplies	3

Table 4.4: Threshold month count to differentiate active and inactive users

4.3 Data clustering

To understand different types of reviewers, we apply clustering to group reviewers with similar review patterns together. The centroid of each cluster represents the general behavior of all the members in the cluster. As stated in Chapter 2, we use *X*-means over the popular *K*-means clustering because of the two main limitations of *K*-means: (1) *K*-means scales computationally poorly, and (2) the number of clusters *K* has to be supplied by user (Pelleg *et al*., 2000). As we are working with large Amazon product review data set we don't want to provide the explicit number of clusters. *X*-means clustering generates as many clusters as necessary, which will help us to understand different types clusters representing different types of reviewers. We will talk about features of Amazon product review data set that are used for clustering in Section 4.3.1.

4.3.1 Feature set selection

Feature set of reviewer is a set of attributes used to describe reviewer in Amazon product review data. As stated in Section 3.1.2 in Chapter 2, Amazon product review dataset contains a number of attributes related to review, reviewer and product. To attain our goal of modeling reviewers, we synthesize these attributes to describe each reviewer and create a list of features referred as feature set. Each review in Amazon product review data contains attributes listed in Table 4.5.

Attribute	Explanation

reviewerID	ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin*	ID of the product, e.g. 0000013714
reviewerName*	Name of the reviewer
helpful	Helpfulness rating of the review, e.g. 2/3
reviewText	Text of the review
overall	rating of the product
summary	summary of the review
unixReviewTime	time of the review (unix time)
reviewTime*	Time of the review (raw)

Table 4.5: Attributes in Amazon product review data (attributes with * are not used for our research)

As listed in Table 4.5, *helpful* attribute of a review is determined by the number of votes received by the review from other customers. As covered in Chapter 3, helpfulness of a review is 2/3, if 2 customers up-voted the review and 1 customer down-voted the review. *Overall* is the rating, ranging from 1 to 5 that the reviewer themselves assign to the product being reviewed. Some of the attributes in Table 4.5 such as *asin*, and *reviewerName* do not help in pattern recognition therefore we choose to ignore them. Also, we derive review posted timestamp from *unixReviewTime* and ignore *reviewTime*. The attributes from Table 4.5 are used to create a feature set to describe each reviewer. Most of this synthesis process is done using Python scripts to create the feature set listed in Table 4.6.

Feature set	**Explanation**
Reviewer ID	ID of the reviewer, e.g. A2SUAM1J3GNN3B
Total review count	Total number of reviews written by the reviewer has, e.g. 5
Average helpfulness	Average of helpfulness of the reviews written by the reviewer, e.g. 1/3
Average review length	Average of the length of reviews written by the reviewer, e.g. 54.06
Average overall	Average of all the rating of the product rated by the reviewer
Total active month	Total number of months the reviewer has been writing reviews

Table 4.6: Feature set of each reviewer

While creating the feature set for clustering, we try to synthesize as many features as possible. Although, in case of multidimensional data, features are selected in such a way that they can cover all the possible clusters in the data because different features have different power on differentiating different clusters (Cai *et al*., 2010). In our case, we use

6 out of 9 different attributes as listed in Table 4.5 to describe reviews, we were able to synthesize 6 different features to describe reviewers.

4.3.2 Data clustering result

After synthesizing the feature set, the reviewers are automatically divided into clusters by using *X*-means clustering. The two-step process of first creating feature set and second performing *X*-means clustering on the featured dataset are repeated on each product category listed in Table 4.1. The number of clusters found is different for different product category as seen in Table 4.7.

Product categories	Number of clusters
Books	3
Electronics	3
Cell Phones and accessories	2
Grocery and gourmet food	4
Health and personal care	4
Office product	3
Baby	4
Beauty	4
Pet supplies	4

Table 4.7: Number of clusters for each category

The centroid of each cluster is a representative of the respective cluster. The cluster centroid consist of same features that were used to describe each reviewer, as listed in Table 4.6. Based on the feature set of the cluster centroid, the respective cluster is differentiated as either expert or novice or conscientious, and so on. The process of analyzing cluster centroid to label each clusters is explained in Section 4.3.2.

4.3.3 Data cluster analysis

We perform cluster analysis of all nine categories in three steps. First, we take Books as the first category, diagnose the behavior of each cluster in this category by observing its centroid then use *t*-test and graphs to differentiate these clusters based on their nature. Second, we repeat the similar process to differentiate clusters into different types in

remaining eight categories (e.g., Electronics, Cell Phones and accessories, Health and personal care, Grocery and gourmet food, Office product, Baby, Beauty, and Pet supplies) highlighting the differences and similarities of clusters in different categories. Third, we draw conclusions based on the observations we made in first and second step.

4.3.3.1 Data cluster analysis- Step 1 (Analysis process)

The centroid of a cluster is the middle of the cluster and represents the average across all the points in the cluster. A cluster's centroid gives us an idea about the general nature of the cluster based on the values of the centroid's features. We analyze the features of each centroid in Books category and use this to differentiate the clusters into expert or novice or any other type as necessary.

Books

From Table 4.7, we can observe that, in the Books category, reviewers are clustered into 3 different clusters. Table 4.8 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	18.711	771.020	3.484	9.268	0.644	93438 (30%)
C2	**25.151**	**837.275**	4.530	**9.946**	**0.854**	129164 (41%)
C3	15.954	491.831	**4.627**	7.977	0.659	92721 (29%)

Table 4.8: Cluster centroid feature values for "Books" category. Bolded values indicate highest values.

The cluster centroids in Table 4.8 tell us the general behavior of each cluster as a whole and we can make the following quick observations:

- Reviewers in C2 have the highest values for all attributes except average overall, which it comes in a close second.
- Reviewers in C3 have the highest average overall among the three clusters.

- Reviewers in C3 have the least total review count, average review length and total active month among the three clusters.
- Reviewers in C1 have the least average overall and average helpfulness among the three clusters.

We perform *t*-test analysis to find if the quick observations we made about each feature are statistically significant.

First, Table 4.9 shows additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values. Standard deviation measures how dispersed the numbers are within the range of minimum and maximum value. Higher standard deviation indicates that the data points are spread far from the mean value. For example standard deviation of total review count in C2 is 141.929, which is very high compared to other clusters. Figure 5(b) of total review count with respect to active month for C2 shows that there are a few data points that are very far from the mean value where most of the other data points are located. So standard deviation is a good indicator of knowing if majority of the reviewers within the cluster strictly follow same trend or vary widely.

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	18.711	5	4050	45.041
C2	25.151	5	35625	141.929
C3	15.954	5	1316	22.483
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	771.020	81.909	29980.833	777.325
C2	837.275	76.800	18490.833	783.724
C3	491.831	67.125	20928.333	504.424
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	3.392	1	4.204	0.540
C2	4.523	3.250	5	0.369
C3	4.585	4.038	5	0.292
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	9.548	5	132	8.720
C2	9.935	5	132	10.099
C3	7.863	5	120	5.111
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	0.644	0	1	0.127
C2	0.854	0.700	1	0.065
C3	0.659	0	0.770	0.100

Table 4.9: Statistics of feature set in different clusters for "Books" category

Table 4.10 below shows the *t*-test results for the three clusters in the Books category for each feature.

	***p*-value**				
Pairs	**Total review count**	**Average review length**	**Average overall**	**Total active month**	**Average helpfulness**
C1 vs. C2	**1.02878E-22**	**8.76924E-06**	**0**	**1.45806E-09**	**0**
C1 vs. C3	**2.9947E-52**	**0**	**0**	**2.1523E-183**	**4.566E-105**
C2 vs. C3	**1.02653E-86**	**0**	**7.0889E-116**	**1.4871E-225**	**0**

Table 4.10: *t*-Test result for clusters in "Books" category. Bolded values (all) are statistically significant, $p < 0.05$.

From the *p*-value of *t*-test on all the features as displayed in Table 4.10, we observe that $p < 0.05$ which is strong evidence against null hypothesis so we accept alternative hypothesis- which states that each pair of mean values of each feature in 2 different

clusters are not equal. So based on the alternative hypothesis we can make the following conclusions from Table 4.9:

- Mean of total review count in C2 is greater than C3, which in turn is greater than C1. This can be observed visually in Figures 4.5(a), 4.5(b) and 4.5(c) that show the distribution of total review count over active month for C1, C2 and C3 respectively.
- Mean of average review length of C2 is greater than C1, which in turn is greater than C3. This can be observed visually in Figure 4.6(a), 4.6(b) and 4.6(c) that show the distribution of average review length over active month for C1, C2 and C3 respectively.
- Mean of average overall of C3 is greater than C2, which in turn is greater than C1. This can be observed visually in Figures 4.7(a), 4.7(b) and 4.7(c) that show the distribution of average overall over active month for C1, C2 and C3 respectively.
- Mean of total active month of C2 is greater than C1, which in turn is greater than C3. This can be observed visually in Figures 4.8(a), 4.8(b) and 4.8(c) that show the distribution of total active month over total review count for C1, C2 and C3 respectively.
- Mean of average helpfulness of C2 is greater than C3, which in turn is greater than C1. This can be observed visually in Figures 4.9(a), 4.9(b) and 4.9(c) that show the distribution of average helpfulness over active month for C1, C2 and C3 respectively.

Figures 5-9 show comparative distribution of each feature set in different clusters. They give visual acknowledgement to the above observations.

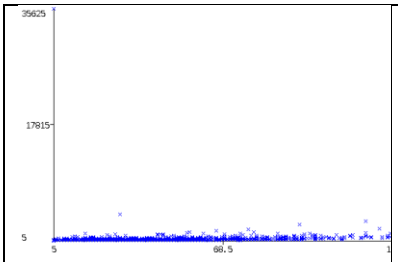

Figure 4.5(a): Graph of total review count with respect to active month for C1

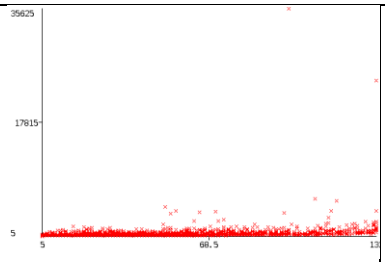

Figure 4.5(b): Graph of total review count with respect to active month for C2

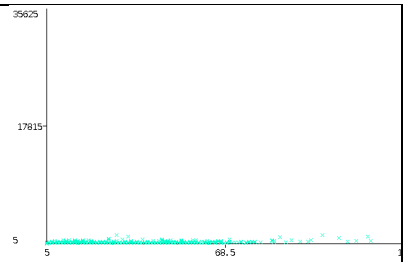

Figure 4.5(c): Graph of total review count with respect to active month for C3

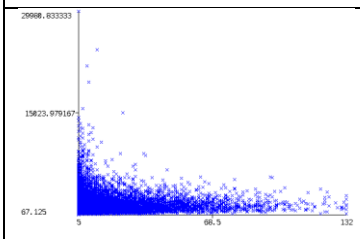

Figure 4.6(a): Graph of average review length with respect to active month for C1

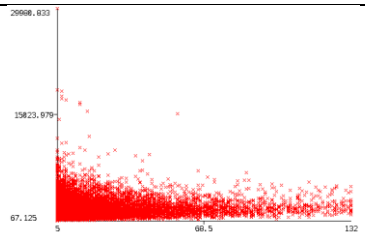

Figure 4.6(b): Graph of average review length with respect to active month for C2

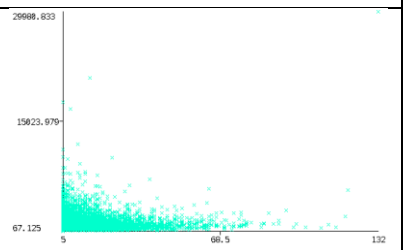

Figure 4.6(c): Graph of average review length with respect to active month for C3

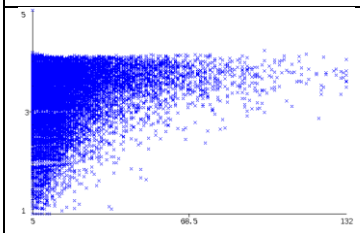

Figure 4.7(a): Graph of average overall with respect to active month for C1

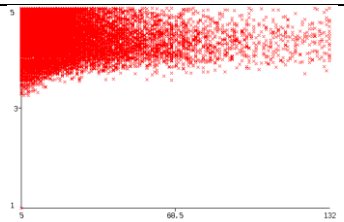

Figure 4.7(b): Graph of average overall with respect to active month for C2

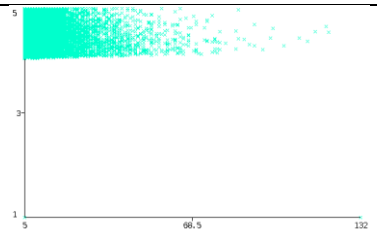

Figure4. 7(c): Graph of average overall with respect to active month for C3

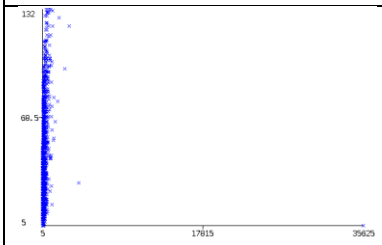

Figure 4.8(a): Graph of total active month with respect to total review count for C1

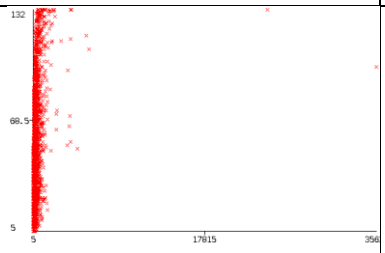

Figure 4.8(b): Graph of total active month with respect to total review count for C2

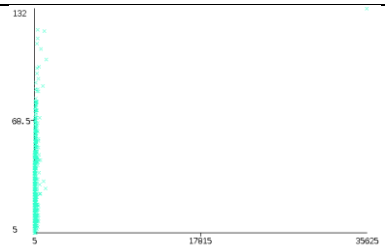

Figure 4.8(c): Graph of total active month with respect to total review count for C3

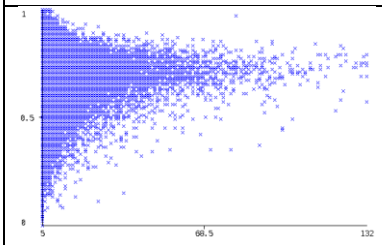

Figure 4.9(a): Graph of average

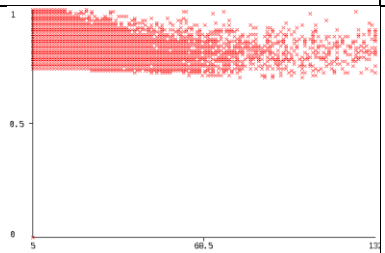

Figure 4.9(b): Graph of average

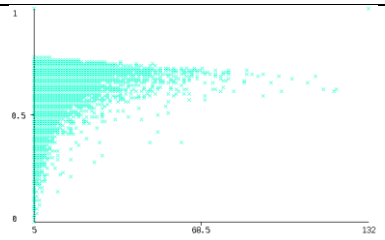

Figure 4.9(c): Graph of average

helpfulness with respect to active month for C1	helpfulness with respect to active month for C2	helpfulness with respect to active month for C3

Based on the observations we made from Table 4.9, which are supported by Figures 4.5-4.9 we will name the three clusters according to their nature defined by the values of their feature set. For example, features such as review count and review length measure the *devotion* of a reviewer towards online reviewing. Furthermore, a feature like overall measures the *worthiness* of the product to a reviewer perceived by the reviewer themself. Together, these features are entirely dependent on the reviewer's devotion and perception. We refer to these features as *internal* features since reviewer has total control over these features. On the other hand, as stated in Chapter 3, a feature like helpfulness is a measure of how useful/helpful the review is as perceived by other users. Helpfulness of a review measures the worthiness of the review to other users and is referred as an *external* feature. An *external* feature is mostly unbiased, as a reviewer has no direct control over it—except for writing a good or bad review. An *external* feature helps to provide unbiased quantification of the quality of review to some degree.

Based on the *external* and *internal* features we name the clusters accordingly as described below.

- Cluster C2 represents reviewers who have highest review count, longest review length, most helpful reviews and are active for the longest period of time. Based on their high interest for reviewing and the helpfulness of their reviews to other users, we refer to them as the *expert* cluster.
- Cluster C3 represents reviewers who have least review count, shortest review length and are active for least amount of time. From this, we can say that these reviewers are not very interested in reviewing and we refer to them as the *novice* cluster. However

they tend to give highest overall to the products they review and their reviews are less helpful than *expert* cluster C2 but more helpful than C1.

- Cluster C1 represents reviewers who have intermediate review count, intermediate review length, least helpfulness and are active for intermediate amount of time. Intermediate values imply the values are greater than *novice* cluster and less than *expert* cluster. From this tendency we can say that these reviewers are interested and diligent in writing reviews but they lack the idea of writing helpful reviews and we refer to them as the *conscientious* cluster.

We repeat the above process for remaining eight categories and find if the clusters are similar or different than Books category.

4.3.3.2 Data cluster analysis- Step 2 (Cluster types)

In this section we perform the same cluster analysis in five other categories e.g., Electronics, Cell phones and accessories, Health and personal care, Grocery and gourmet food, Office product, Baby, Beauty, and Pet supplies. We also discuss the similarities and differences of cluster nature in different categories.

Electronics

From Table 4.7, we know that there are 3 different types of clusters in electronics category. Table 4.11 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	7.502	**746.024**	2.968	5.798	0.714	33127 (27%)
C2	**9.548**	735.983	**4.3908**	**6.268**	**0.852**	70495(42%)
C3	9.197	512.683	4.349	5.856	0.600	42535(31%)

Table 4.11: Cluster centroid feature values for "Electronics" category. Bolded values indicate highest values.

The cluster centroids in Table 4.11 tell us the general behavior of each cluster as a whole.

Additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values is in Appendix, Section B. We perform *t*-test analysis to find if the values of each feature set in Table 4.11 are statistically significant.

	***p*-value**				
Pairs	**Total review count**	**Average review length**	**Average overall**	**Total active month**	**Average helpfulness**
C1 vs. C2	**1.751E-171**	0.038	**0**	**6.538E-50**	**0**
C1 vs. C3	**1.406E-186**	**0**	**0**	0.039	**0**
C2 vs. C3	**8.169E-06**	**0**	**3.385E-34**	**1.123E-36**	**0**

Table 4.12: t-Test result for clusters in "Electronics" category. Bolded values are statistically significant, p < 0.05.

In Table 4.12, we see that all of the features are statistically significant except for average review length between C1 and C2 and total active month between C1 and C3. Since cluster C1 has the highest average review length observed from Table 4.11, and average review length of C2 is statistically insignificant with respect to C1, we can say that C1 and C2 both have longer average review length compared to C3. Similarly mean of total active month of C1 and C3 are statistically insignificant so their means are equal and both are less than C2.

Based on the *external* and *internal* features, there are three types of clusters in Electronics which are very similar to Books: 1) C2 referred as *expert* cluster, 2) C1 referred as *novice* cluster, and 3) C3 referred as *conscientious* cluster. Expert cluster in both categories represent most experienced reviewers in terms of active months and review count. Similarly in both categories higher review count, active month and overall correspond to higher helpfulness.

From Tables 4.8 and 4.11, we have to note that there are similarities between the *expert* clusters of Books and Electronics respectively. Experts in both the categories have the highest helpfulness which may be largely because they 1) review frequently i.e., have

highest review count, 2) have been active for the longest period of time i.e., have highest active month, and 3) share satisfied positive experiences with product i.e., have high overall.

Cellphones and accessories

From Table 4.7, we know that there are 2 types of clusters in cellphones and accessories category. Table 4.13 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	6.784	**562.058**	3.210	4.929	0.678	7740
C2	**7.851**	537.654	**4.392**	**5.142**	**0.768**	13946

Table 4.13: Cluster centroid feature values for "Cell Phones and accessories" category. Bolded values indicate highest values.

The cluster centroids in Table 4.13 tell us the general behavior of each cluster as a whole. Additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values is in the Appendix, Section B.

Table 4.14 displays *p*-value from *t*-test of each attribute in two clusters with a sample size of about 7000.

	p-value				
Pairs	**Total review count**	**Average review length**	**Average overall**	**Total active month**	**Average helpfulness**
C1 vs. C2	**1.804E-43**	**1.477E-13**	**0**	**7.438E-09**	**1.616E-37**

Table 4.14: t-Test result for clusters in "Cell Phones and accessories" category. Bolded values (all) are statistically significant, p < 0.05

From Table 4.14, we see that all of the features are statistically significant since $p < 0.05$. Based on the *external* and *internal* features, there are only two types of clusters: 1) C2 referred as *expert* cluster, and 2) C1 referred as *conscientious* cluster. Note that, however, there is not a third cluster (e.g., a novice cluster as in Books and Electronics). There are fewer reviews per product (~27 reviews per product) in Cell phones and

accessories compared to similar product like Electronics (~38 reviews per product). This is mainly because most people with less expertise tend to buy cellphones and its accessories in-store after discussing the specification of the product with salespersons. Naïve users who otherwise would have been novice reviewers, may find it more easy and reliable to make purchases on relatively high investment products such as cellphones. Furthermore, the ease of buying cellphone at local wireless carrier's store is less time consuming compared to buying the same product online. To make in-store purchase of cellphone a person can just walk into a store, pick a cellphone, set up a plan, and its ready to use. Further, local carrier stores offer various attractive money-saving schemes on data plan bundled together with new cellphones which is not offered if one decides to buy the cellphone online. Additionally, they help their customers to transfer their data plan from old phone to new phone. Hence this ease and convenience attracts most of the novice reviewers who end up making in-store purchase when it come to cellphones and accessories. Even if some novice reviewers may be buying cell phones and accessories online, they are too few in number to warrant a cluster.

Briefly, the characteristics of expert and conscientious clusters are similar to those found in Books and Electronics. Experts represent experienced reviewers in terms of active month and review count. Higher review count, active month and overall correspond to higher helpfulness.

Similar to Books and Electronics, the expert cluster in Cell phones and accessories has the highest helpfulness which can be attributed to the highest overall, highest review count and highest active length as seen in Table 4.13.

Hence we can conclude that:

Conclusion 4a: High overall, review count and active month leads to helpful reviews for Books, Electronics, and Cellphones and accessories.

Health and personal care

From Table 4.7, we know that there are 4 types of clusters in Health and personal care category. Table 4.15 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	6.234	556.852	4.569	4.421	0.632	20284
C2	5.368	**598.995**	3.090	4.221	**0.610**	9719
C3	**6.684**	352.357	**4.664**	4.358	0.238	25107
C4	6.342	415.873	3.468	**4.477**	0.247	15559

Table 4.15: Cluster centroid feature values for "Health and personal care" category. Bolded values indicate highest values.

The cluster centroids in Table 4.15 tell us the general behavior of each cluster as a whole. Additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values is in the Appendix, Section B.

Table 4.16 displays *p*-value from *t*-test of each attribute in four clusters with a sample size of about 7000.

	p-value				
Pairs	**Total review count**	**Average review length**	**Average overall**	**Total active month**	**Average helpfulness**
C1 vs. C2	**4.974E-31**	**1.477E-13**	**0**	**7.438E-09**	**1.616E-37**
C1 vs. C3	**1.702E-10**	**0**	**4.086E-173**	0.185	**0**
C1 vs. C4	0.173	**1.006E-246**	**0**	0.119	**0**
C2 vs. C3	**1.688E-77**	**0**	**0**	**2.238E-06**	**0**
C2 vs. C4	**1.913E-38**	**0**	**0**	**2.179E-06**	**0**
C3 vs. C4	**5.288E-07**	**0**	**0**	**2.417E-06**	**0**

Table 4.16: *t*-Test result for clusters in "Health and personal care" category. Bolded values are statistically significant, $p < 0.05$.

From Table 4.16, we see that all of the features are statistically significant since $p < 0.05$ except total review length and total active month in C1 and C4, and total active month in

C1 and C3. Based on the *external* and *internal* features we interpret the clusters accordingly as described below.

- Cluster C1 represents reviewers who have greater review count, greater review length, have more helpful reviews, have high overall and are active for long period of time. Based on their high interest for reviewing and the helpfulness of their reviews to other users, they are expert cluster. Since they tend to write more reviews compared to cluster C2 which is also an expert cluster, we may refer to C1 as the *frequent expert*. Also they tend to write positive reviews and we may refer to them as *positive expert*.
- Cluster C2 represents reviewers, who have lowest review count, longest review length, have lowest overall and are active for least period of time. From lowest active month and review count, we can say that these reviewers are new reviewers who haven't written many reviews but their reviews are long and detailed, which make them most helpful. So we refer to this cluster as the *non-frequent expert.* They also have the lowest overall which means they tend to write their negative experiences with the product and we may refer to them as *negative expert*.
- Cluster C3 is referred as the *conscientious* cluster.
- Cluster C4 is referred as the *novice* cluster.

In the first three categories analyzed, we observed that higher review count, active month and overall correspond to higher helpfulness.

We observed a very unique cluster referred as *non-frequent negative expert*. Experts we have known so far—e.g., from Books, Electronics, and Cellphones and accessories—are reviewers with more experiences who usually write positive reviews

and have been active for a long time. But these *non-frequent negative experts* are different. They write negative reviews and yet they have highest helpfulness and also have fewer number of reviews. This may be related to Health and personal care product being very different from other products we have covered thus far. Health and personal care are sensitive products as its consumption effects consumers' health directly so users are very careful before purchasing these products. That is, unlike previous categories such as Cell phones, Books and Electronics, users find negative reviews helpful because they want to be well informed on both positive as well as negative effects of the product. Users tend to read through both positive and negative reviews and find detailed and well described reviews more helpful.

One of the important observations that can be drawn from Table 4.15 is that for the two expert clusters that we identify—(1) *non-frequent negative expert* (C2) and (2) *frequent positive expert* (C1)—the average review text length is highest and second highest, respectively, compared to other clusters. These values are statistically significant as seen in Table 4.16. The average helpfulness are highest and second highest for *non-frequent negative expert* (C2) and 2) *frequent positive expert* (C1) respectively.

Hence we draw the following conclusion:

Conclusion 4b: Longer reviews lead to helpful reviews for Health and personal care products.

Grocery and gourmet food

From Table 4.7, we know that there are 4 types of clusters in Grocery and gourmet food category. Table 4.20 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	**8.862**	428.211	3.641	**5.944**	0.186	6756
C2	7.359	473.124	4.596	4.808	**0.654**	9298
C3	7.158	318.656	**4.707**	4.583	0.234	13521
C4	7.245	**491.464**	3.24	5.085	0.515	5136

Table 4.17: Cluster centroid feature values for "Grocery and gourmet food" category. Bolded values indicate highest values.

The cluster centroids in Table 4.17 tell us the general behavior of each cluster as a whole. Additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values is in the Appendix, Section B.

Table 4.18 displays *p*-value from *t*-test of each attribute in four clusters with a sample size of about 5000.

	***p*-value**				
P-Value	**Total Review Count**	**Average Review Length**	**Average Overall (rating)**	**Total Active Month**	**Average Helpfulness**
C1 vs. C2	**7.18361E-14**	**3.93857E-17**	**0**	**7.44997E-48**	**0**
C1 vs. C3	**1.90958E-26**	**1.3657E-144**	**0**	**7.42498E-82**	**2.0396E-154**
C1 vs. C4	**2.02067E-16**	**3.09717E-23**	**1.4369E-270**	**1.02828E-22**	**0**
C2 vs. C3	0.220371617	**4.0523E-274**	**8.1817E-126**	**1.88921E-05**	**0**
C2 vs. C4	0.566135001	**0.004018174**	**0**	**0.000159342**	**0**
C3 vs. C4	0.582984919	**5.6938E-204**	**0**	**1.9329E-14**	**0**

Table 4.18: *t*-Test result for clusters in "Grocery and gourmet food" category. Bolded values are statistically significant, $p < 0.05$.

From Table 4.18, we see that all of the features are statistically significant since $p < 0.05$ except total review length C2 and C3, C2 and C4, and C3 and C4. This means total review count of C2, C3 and C4 are equal to each other and all are less than C1. Based on the *external* and *internal* features we interpret the clusters accordingly as described below.

- Cluster C2 represents reviewers who have greater review count, greater review length, have most helpful reviews, have high overall and are active for long period of time. Based on their high interest for reviewing and the helpfulness of their reviews to

other users, they are expert cluster. Since they tend to give positive rating compared to cluster C4 which is also an expert cluster, we refer to C2 as the *positive expert*.

- Cluster C4 represents reviewers, who have greater review count, greatest review length, have more helpful reviews, have least overall and are active for long period of time. From least overall and highest review length, we can say that these reviewers write detailed review for products that they are not satisfied with. These reviews are found to be very helpful to other uses so we refer to this cluster as the *negative expert*. Long and detailed reviews may be the reason of why these reviews are most helpful.
- Cluster C1 is referred as the *conscientious* cluster.
- Cluster C3 is referred as the *novice* cluster.

Similar to Health and personal care, we observe that longer the review length more helpful the reviews are. So for Grocery and gourmet food, a detailed well described reviews are found to be more helpful irrespective of overall rating.

Again, while we observed that higher overall leads to helpful reviews in the first three categories—Books, Electronics, and Cell phones and accessories, this is not true for Grocery and gourmet food as we observe that both higher and lower overall—i.e., C2 and C4 respectively—lead to helpful reviews. This suggests that reviews for Grocery and gourmet foods are helpful for both positively and negatively rated products. An example of a helpful negative review is a review that informs that a food item causes a specific type of allergy in babies. This review could help consumers to make informed purchase decision which may be either to buy the food for adults or not to buy the food for babies. Despite the bad review, the review turned out to be very helpful.

Considering C4 as *Negative experts,* they are similar to the one we observed in Health and personal care. Like Health and personal care, Grocery and gourmet food are sensitive products as they directly affect consumers' health. So users want to read through all kinds of reviews both positive and negative and make a well-informed decision whether to buy the product or not. We observe from Table 4.17 that for expert clusters identified—(1) *negative expert* (C4) and (2) *positive expert* (C2), the average review text length is the highest and second highest respectively compared to other clusters. These values are statistically significant as seen in Table 4.18. The average helpfulness are highest and second highest for *positive expert* (C2) and *negative expert* (C4) respectively. Hence we can established that, the common feature between positive and negative expert is the practice of writing long, well described reviews which is the reason behind their helpfulness. So we can infer that users find it helpful when they read the details of product whether positive or negative when buying Grocery and gourmet foods. Similar to Health and personal care products we conclude that:

Conclusion 4c: Longer reviews leads to helpful reviews for Grocery and gourmet food products.

Unlike Health and Personal care products, the reviewing frequencies for both positive and negative experts in Grocery and gourmet food are similar. This may be because food items are one of the basic requirements of humans and they seem to review these products pretty well without much practice or experience. From Table 4.17, we know that *negative expert* (C4) and *positive expert* (C2) both have lower review count compared to other clusters. Table 4.18 supports that the reviewing frequency of these two clusters are not statistically significant or in other words, both the clusters have same low

reviewing frequency. This is very unique to Grocery and gourmet food as experts in other categories usually have high reviewing frequency or active month. Hence we draw following conclusion:

Conclusion 4d: Expertise for reviewing Grocery and gourmet food does not necessarily come with practice.

Office products

From Table 4.7, we know that there are 3 types of clusters in Office products category. Table 4.19 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	5.545	**764.199**	**4.398**	**4.657**	**0.661**	6766
C2	4.266	717.387	2.652	3.765	0.539	3527
C3	**6.080**	442.537	4.368	4.448	0.213	12806

Table 4.19: Cluster centroid feature values for "Office product" category. Bolded values indicate highest values.

The cluster centroids in Table 4.19 tell us the general behavior of each cluster as a whole. Additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values is in the Appendix, Section B.

Table 4.20 displays *p*-value from *t*-test of each attribute in three clusters with a sample size of about 3000.

	***p*-value**				
Pairs	**Total review count**	**Average review length**	**Average overall**	**Total active month**	**Average helpfulness**
C1 vs. C2	**1.72893E-59**	**0.0002E-5**	**0**	**2.72097E-45**	**5.5505E-195**
C1 vs. C3	**1.73745E-10**	**8.5285E-257**	**7.61643E-05**	**3.35378E-05**	**0**
C2 vs. C3	**7.4868E-168**	**1.4426E-147**	**0**	**3.5642E-102**	**0**

Table 4.20: *t*-Test result for clusters in "Office product" category. Bolded values (all) are statistically significant, $p < 0.05$.

From Table 4.20, we see that all of the features are statistically significant since $p < 0.05$. Based on the *external* and *internal* features there are again three types of clusters: (1) C1

referred as *expert* cluster (2) C2 referred as *novice* cluster, and (3) C3 referred as *conscientious* cluster. Also, the characteristics of three clusters in office products are similar to Books, Electronics, and Cell phones and accessories because high overall leads to more helpfulness.

Expert cluster C1 has the highest overall and helpfulness as seen in Table 4.19.

Hence we can conclude that:

Conclusion 4e: High overall leads to helpful reviews for Office product.

Baby

From Table 4.7, we know that there are 4 types of clusters in Baby category. Table 4.21 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	8.193	648.17	3.284	5.272	0.327	1139
C2	**11.195**	679.791	4.394	**6.144**	0.589	1872
C3	9.602	432.561	**4.413**	4.939	0.137	2008
C4	8.41	**1035.95**	3.579	5.688	**0.68**	1014

Table 4.21: Cluster centroid feature values for "Baby" category. Bolded values indicate highest values.

The cluster centroids in Table 4.21 tell us the general behavior of each cluster as a whole. Additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values is in the Appendix, Section B.

Table 4.22 displays *p*-value from *t*-test of each attribute in three clusters with a sample size of about 1000.

	***p*-value**

Pairs	Total review count	Average review length	Average overall	Total active month	Average helpfulness
C1 vs. C2	**6.90769E-38**	**0.024563075**	**0**	**2.87749E-22**	**4.11864E-45**
C1 vs. C3	**2.28494E-14**	**3.22345E-60**	**0**	**1.05307E-07**	**2.0356E-284**
C1 vs. C4	**0.343998611**	**2.39739E-59**	**3.3248E-184**	**3.89249E-05**	**0**
C2 vs. C3	**2.56099E-11**	**4.1414E-128**	**0.096658381**	**8.37046E-51**	**0**
C2 vs. C4	**9.80313E-28**	**3.7032E-55**	**4.552E-100**	**3.24056E-05**	**0**
C3 vs. C4	**1.92226E-08**	**1.5035E-139**	**2.0361E-104**	**3.81957E-17**	**0**

Table 4.22: *t*-Test result for clusters in "Baby" category. Bolded values (all) are statistically significant, $p < 0.05$.

From Table 4.22, we see that all of the features are statistically significant since $p < 0.05$. Based on the *external* and *internal* features there are again four types of clusters: (1) C1 referred as *novice* cluster, (2) C2 referred as *frequent positive expert* cluster, (3) C3 referred as *conscientious* cluster, and (4) C4 referred as *non-frequent negative expert* cluster. These clusters are similar to Health and personal care, and Grocery and gourmet food as they have two different types of experts.

Baby product includes foods, milk bottles, diapers, wipers, etc. designed specifically for babies. New parents who are the biggest buyers of these products are very sensitive with regard to baby's health, nutrition, and comfort so they buy the best of what is available in the market and try to avoid products with any negative consequences. The risk involved in buying negatively reviewed product is very high for baby products so they find negative reviews very helpful as it informs them of bad experiences.

From Table 4.21, we see that for two expert clusters—(1) *non-frequent negative expert* (C4) and (2) *frequent positive expert* (C2), the review text length is highest and second highest respectively compared to other clusters. These values are statistically significant as seen in Table 4.22. Subsequently, the average helpfulness are highest and second highest for (1) *non-frequent negative expert* (C4) and (2) *frequent positive expert*

(C2), respectively. Hence, similar to Health and personal care, and Grocery and gourmet food we can draw following conclusion from this.

Conclusion 4f: Longer reviews lead to helpful reviews for Baby products.

Beauty

From Table 4.7, we know that there are 4 types of clusters in Beauty category. Table 4.23 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	6.657	523.868	4.595	4.327	**0.644**	12138
C2	6.497	**569.676**	3.294	**4.376**	0.605	6728
C3	**7.149**	332.453	**4.654**	4.137	0.233	14803
C4	6.954	395.392	3.452	3.452	0.239	9116

Table 4.23: Cluster centroid feature values for "Beauty" category. Bolded values indicate highest values.

The cluster centroids in Table 4.23 tell us the general behavior of each cluster as a whole. Additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values is in the Appendix, Section B.

Table 4.24 displays *p*-value from *t*-test of each attribute in four clusters with a sample size of about 6000.

	***p*-value**				
Pairs	**Total review count**	**Average review length**	**Average overall**	**Total active month**	**Average helpfulness**
C1 vs. C2	0.155039331	**9.11616E-12**	**0**	0.343775814	**2.50029E-83**
C1 vs. C3	**3.31059E-05**	**0**	**7.53052E-42**	**7.69996E-09**	**0**
C1 vs. C4	**0.005378707**	**4.4047E-161**	**0**	0.193982005	**0**
C2 vs. C3	**7.04683E-07**	**0**	**0**	**2.87886E-09**	**0**
C2 vs. C4	**0.000125949**	**1.4422E-179**	**2.66856E-94**	**0.043376118**	**0**
C3 vs. C4	0.218105277	**1.30676E-60**	**0**	**1.23826E-05**	**0.003773504**

Table 4.24: *t*-Test result for clusters in "Beauty" category. Bolded values are statistically significant, $p < 0.05$.

From Table 4.24, we see that all of the features are statistically significant since $p < 0.05$ except total review count and total active month in C1 and C2, and total active month in

C1 and C4. Based on the *external* and *internal* features we interpret the clusters accordingly as described below.

- Cluster C1 represents reviewers who have a moderate review count, greater review length, have more helpful reviews, have high overall and are active for long period of time. Based on their high interest for reviewing and the helpfulness of their reviews to other users, they constitute the expert cluster. Since they tend to write more positive reviews compared to cluster C2 (discussed below), which is also an expert cluster, we may refer to C1 as the *positive expert* cluster.
- Cluster C2 represents reviewers who have a moderate review count similar to C1 (*positive expert*), longest review length, have lowest overall and are active for longest period of time similar to C1 (*positive expert*) above. So we refer to this cluster as the *expert.* They have the lowest overall which means they tend to write their negative experiences with the product and we may refer to them as *negative expert*.
- Cluster C3 is referred as the *conscientious* cluster.
- Cluster C4 is referred as the *novice* cluster.

Considering C2 as *negative experts*, they are similar to the one observed on Grocery and gourmet food, Health and personal care, and Baby products. Beauty products such as skin care products, hair products, make up products, personal care products and so on are sensitive products as they directly affect consumers' health. So users want to know about both the positive and negative experiences and make a well-informed decision whether to buy the product or not. One of the important observations that can be drawn from Table 4.23 is that for the two expert clusters that we identify—(1) *positive expert* (C1) and (2) *negative expert* (C2) —the average review text length is the

second highest and highest, respectively, compared to the other two clusters. These values are statistically significant as seen in Table 4.24. The average helpfulness are highest and second highest for *positive expert* (C1) and *negative expert* (C2) respectively.

Hence we draw the following conclusion:

Conclusion 4g: Longer reviews lead to helpful reviews for Beauty products.

Pet supplies

From Table 4.7, we know that there are 4 types of clusters in Pet supplies category. Table 4.25 depicts the values of the feature set of the centroid of each cluster.

Cluster number	Total review count	Average review length	Average overall	Total active month	Average helpfulness	Observations
C1	6.154	625.809	4.512	4.324	**0.633**	7730
C2	5.728	**668.536**	3.161	**4.168**	0.604	4040
C3	**6.476**	381.065	**4.663**	3.999	0.210	11393
C4	6.0813	430.376	3.515	3.964	0.207	6553

Table 4.25: Cluster centroid feature values for "Pet supplies" category. Bolded values indicate highest values.

The cluster centroids in Table 4.25 tell us the general behavior of each cluster as a whole. Additional statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values is in the Appendix, Section B. Table 4.26 displays *p*-value from *t*-test of each attribute in four clusters with a sample size of about 4000.

	p-value				
Pairs	**Total review count**	**Average review length**	**Average overall**	**Total active month**	**Average helpfulness**
C1 vs. C2	**2.72106E-09**	**6.00856E-07**	**0**	**1.2881E-06**	**7.08937E-10**
C1 vs. C3	**0.002969659**	**0**	**6.1912E-153**	**9.33346E-23**	**0**
C1 vs. C4	0.629438083	**9.4263E-177**	**0**	**1.61836E-18**	**0**
C2 vs. C3	**1.35707E-23**	**9.2812E-229**	**0**	**0.000483534**	**0**
C2 vs. C4	**2.88792E-09**	**2.0504E-143**	**2.0981E-241**	0.005547747	**0**
C3 vs. C4	**0.000152592**	**1.06199E-34**	**0**	0.599510205	**0.002495584**

Table 4.26: *t*-Test result for clusters in "Pet supplies" category. Bolded values are statistically significant, $p < 0.05$.

From Table 4.26, we see that all of the features are statistically significant since $p < 0.05$ except total review count in C1 and C4, and total active month in C2 and C4, and C3 and

C4. Based on the *external* and *internal* features there are again four types of clusters: (1) C1 referred as *frequent positive expert* cluster, (2) C2 referred as *non-frequent negative expert* cluster, (3) C3 referred as *conscientious* cluster, and (4) C4 referred as *novice* cluster. These clusters are similar to Health and personal care, and Baby as they have two different types of experts.

Pet supplies consist of products like cat food, dog food, horse food, and related products such as cage for birds, activity tree for cat, playhouse for rabbit and so on. Pet owners are the biggest buyers of these products and are very sensitive to their pet's health and well being so they buy the best of what is available in the market. They try to avoid products with any negative consequences. The risk involved in buying negatively reviewed product is very high for pet supplies so they find negative reviews very helpful as it informs them of any probable bad experiences.

From Table 4.25 we can see that for two expert clusters (1) *frequent positive expert* (C1), and (2) *non-frequent negative expert* (C2), reviews length is second highest and highest respectively. These values are statistically significant as seen in Table 4.26. Subsequently, the average helpfulness are highest and second highest for (1) *frequent positive expert* (C1) and (2) *non-frequent negative expert* (C2), respectively. Hence, similar to Health and personal care, and Grocery and gourmet food, Baby, and Beauty products we can conclude:

Conclusion 4h: Longer reviews lead to helpful reviews for Pet supplies.

4.3.3.3 Data cluster analysis- Step 3 (Summary)

Below we summarize different types of clusters we have observed:

- All the nine categories of products have 3 different types of clusters: (1) *expert*, (2) *novice*, and (3) *conscientious* except for the Cell phones and accessories category. Cell phones and accessories are a very unique product type as they involve comparatively large investment and people with less technical skills are more comfortable buying these products in store after speaking with salesperson. We speculate that, as a result, these buyers who would have made the novice cluster tend to make in-store purchases and are missing from our data.
- There are two types of experts in Health and personal care, Baby, and Pet supplies: (1) *frequent expert*, and (2) *non-frequent expert*. *Frequent expert* is similar to the expert clusters in Books, Electronics, and Cell phones and accessories. *Non-frequent expert* is, on the other hand, rather unique. These reviewers have a very short active period and a small number of review counts which together give an impression of early maturity. It indicates that these reviewers are good at reviewing from the very start. Unlike frequent experts, they don't need time and experience to write helpful reviews. At the same time we have to note that non-frequent experts usually write negative reviews which may make their reviews helpful in these categories. Users are very careful when purchasing products with health concerns like Health and personal care, and Baby products so they find negative reviews more helpful in this regard. See discussion below.
- There are ***two*** types of experts in Health and personal care, Grocery and gourmet food, Baby, Beauty, and Pet supplies: (1) *positive expert*, and (2) *negative expert*. Experts who usually give high overall to products are referred as positive experts as their reviews explain the positive or satisfactory effects of products. Negative experts

are those who share their dissatisfaction with products. Both positive and negative reviews tend to be helpful to users specially when they are related to human/animal health.

Now, we know that that there are different types of clusters in different categories such as *experts*, *frequent experts*, *positive experts* and so on. Existence of different types of experts show that helpfulness of reviews is determined by various other attributes such as positive/negative review, length of review, and so on. Specifically two attributes—overall (rating) and review length—have a prominent effect on helpfulness so we claim two hypotheses and list the categories that abide by each hypothesis.

Hypothesis/claim	Categories list
H1. High overall leads to helpful reviews	Books Electronics Cell Phones and accessories Office product
H2. Longer reviews leads to helpful reviews	Health and personal care Grocery and gourmet food Baby Beauty Pet supplies

Table 4.27: Two hypotheses and categories that meet each hypothesis.

Table 4.27 lists two key attributes, **overall (rating)** and **review length,** that have prominent effect on helpfulness. These two claims are supported by a number of categories. Expert reviewers clusters in categories such as Books, Electronics, Cellphones and accessories, and Office products usually has the highest overall which are highlighted in conclusion 4a and 4e. This shows a direct effect of overall on helpfulness. In other words, higher overall leads to more helpful reviews and lower overall leads to less helpful reviews.

Categories such as Health and personal care, Grocery and gourmet food, Baby, Beauty, and Pet supplies show a direct effect of review length on helpfulness.

Conclusions 4b, 4c, 4f, 4g, and 4h highlight the effect of review length on Health and personal care, Grocery and gourmet food, Baby, Beauty, and Pet supplies, respectively.

Usually when a user evaluates a product review, they may activate a regulatory system that is congruent with the consumption goal. As stated in Chapter 2, there are two separate systems to process product information: one that calls on the *promotion* system to identify useful information for achieving desirable outcomes and the other that calls on the *prevention* system to identify useful information for avoiding undesirable outcomes (Zhang *et al*., 2010). Users with promotion goal are more concerned with advancement and achievement through product consumption. For the three product categories (Books, Electronics, Cellphones and accessories) we can say that user activates promotion consumption goal since people usually want to read or use positively reviewed products when it comes to books, electronics, cellphones and office products. For these types of product, users prefer top-of-the-chart products. A real book lover would not want to miss any best sellers even when there are negative reviews about them. Similarly, a tech-lover would want to try highly recommended electronics, cellphones and accessories. They are not very concerned about the minor flaws of these products, if any, because there is not high risk associated with buying a wrong product. For these products a review with higher overall is perceived as more helpful. Reviews with high overall represent positive product reviews, which provide information about satisfactory experiences with the product, and thus represent opportunities to attain positive outcomes. These reviews are perceived to be more helpful for users with promotion consumption goal.

Conversely, for users who evaluate products associated with prevention consumption goals perceive negative reviews to be more persuasive than positive ones

(Zhang *et al*., 2010). For products such as Health and personal care, Grocery and gourmet food, and Baby, Beauty, and Pet supplies, we observed that negative overall is associated with high helpfulness. With products that are important to consumer in at a personal level such as Health and personal care, Grocery and gourmet food, Baby, Beauty, and Pet supplies, consumers are usually more cautious. They read through negative reviews because they want to avoid negative consequences as much as possible. The penalty associated with any bad reviews if true outweighs the benefits of positive reviews. Therefore, consumers are risk averse for this kind of products. So they also find longer in-depth reviews with negative reviews helpful in addition to positive ones. Hence we can say that user activates prevention consumption goal when it comes to Health and personal care, Grocery and gourmet food, Baby, Beauty, and Pet supplies.

For products associated with promotion consumption goals, positive reviews are more helpful than negative ones whereas for products associated with prevention consumption goals, negative reviews are more helpful than positive ones. We can deduce that perceived helpfulness of a review depends on consumption goal and thus on *product type*. So apart from feature set explained in Section 4.3.1, we should also consider product type when analyzing the helpfulness of reviews.

4.4 Data classification

In Section 4.3 we discussed a three-step clustering and interpretation process: (1) feature extraction, (2) clustering using X-means, and (3) cluster interpretation and analysis in identifying particular classes such as *experts*, *novices*, and *conscientious*. In this section, we create a classifier that is capable of predicting a new reviewer into one of the aforementioned classes based on the feature set of the reviewer.

4.4.1 Data classification using J48

In Section 2.3, we discussed the advantages of using decision tree such as high computational efficiency, easy to understand, and clear rules. We use J48 implementation of C4.5 algorithm for our classification. C4.5 was developed by Ross Quinlan and is used to build decision trees using the concept of Information Entropy (Quinlan, 1993). C4.5 provides computing efficiency, deals with continuous values, handles attributes with missing values, and avoids over fitting by pruning trees after creation (Deepti, *et al*., 2010).

To build a decision tree, a training data set, $S = s_1, s_2 \ldots$ of classified samples is required. Each sample $s = x_1, x_2 \ldots$ is a vector where $x_1, x_2 \ldots$ represent features of the sample. At each node of the tree, J48 chooses one feature of the data that most effectively splits its set of samples into subset belonging in one class or the other. The splitting is performed based on the normalized information gain, which is the difference in entropy, that results from choosing an attribute (Kumar and Rathee, 2011). The attribute with the highest normalized information gain is chosen to make the decision. This process then recurs on the smaller subtrees. The decision is grown using depth-first strategy.

In our case, the training data is a set $S = s_1, s_2 \ldots$ of reviewers already classified into one of the classes such as expert, novice, or conscientious. Each reviewer $s = x_1, x_2 \ldots x_7$ is a vector where $x_1, x_2 \ldots x_7$ represent nine features of the reviewer listed in Table 4.6. We use the data as training set and build a J48 pruned decision tree that will be able to classify new reviewer into one of the classes based on their features. As recommended by WEKA, we use default value of 0.25 pruning confidence, 3 folds for

reduced error pruning, and minimum of 2 instances per leaf for this tree(Bouckaert *et al*., 2013).

For the Books category, its decision tree was trained with 315,323 instances of classified reviewers. The pruned J48 tree obtained has total of 705 nodes, also referred as tree size and contains 353 leaves. We analyzed the J48 pruned decision tree to find which feature is used to split the data at each level.

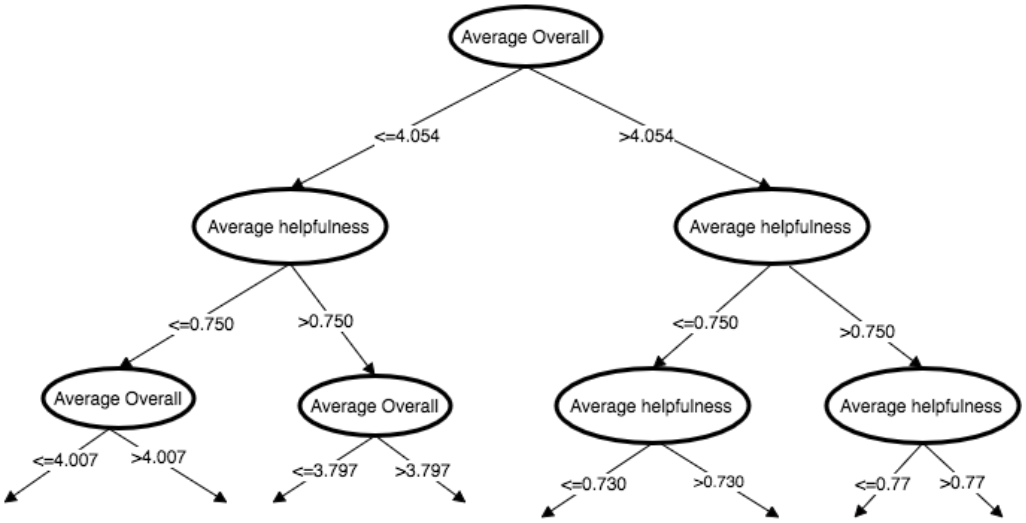

Figure 4.10: First three level of J48 pruned tree of the "Books" category.

Figure 4.3 shows that average overall has the highest normalized information gain and it is used to split the data at root level. At level 2, the factor average helpfulness is used for splitting which means the information gain of average helpfulness is highest at this level. The fact that average overall (root node) has the highest normalized information gain indicates that average overall is the most important feature and its value has the highest weightage in the classification of a new reviewer.

Similarly for Electronics, and Cell phones and accessories, *average overall* has the highest normalized information gain and is the root node in respective decision trees. The decision trees for all the categories are presented in the Appendix, Section C.

The summary of the decision tree analyses for all the categories is presented in Table 4.28.

Category	Number of training instances	Number of leaves	Tree size (# internal nodes + # leaves)
Books	315323	353	705
Electronics	146157	204	407
Cell Phones and accessories	21686	52	103
Health and personal care	70669	79	157
Grocery and gourmet food	34711	121	241
Office product	23099	55	109
Baby	6033	102	203
Beauty	42785	53	105
Pet supplies	29716	48	95

Table 4.28: Summary of decision tree classifiers for all categories.

We now summarize J48 decision tree for each of the nine product categories. For each decision tree, Table 4.29 provides the list of features used in decision nodes at the tree's top three level.

Categories list	Features at Level 1	Features at Level 2	Features at Level 3
Books	average overall	average helpfulness	average overall, and average helpfulness
Electronics	average overall	average overall, average helpfulness	average helpfulness
Cell Phones and accessories	average overall	average overall	average helpfulness
Health and personal care	average helpfulness	average overall	average helpfulness
Grocery and gourmet food	average helpfulness	average overall	average overall, and average helpfulness
Office product	average helpfulness	average overall	average overall, and average helpfulness
Baby	average helpfulness	average overall, average helpfulness	average overall, average helpfulness, and total active month
Beauty	average helpfulness	average overall	average helpfulness, and average overall
Pet supplies	average helpfulness	average overall	average helpfulness, and average overall

Table 4.29: Features used in top 3 level of decision nodes

From Table 4.29 we see that different features such as average overall, average helpfulness, and total active month are used in different levels of a decision tree. On

expanding such a decision tree, we could also observe other features like average review length and total review count as we go deeper in the tree, closer to leaves.

Observing the root node of decision trees for each of the nine product categories in Table 4.29 we can say that average overall and average helpfulness are the two most important features for the classification of reviewers. Specifically, Table 4.30 displays the most important feature and the respective product category and their product consumption goal.

Distinguishing feature	Categories list
Average overall	Books (Promotion consumption goal) Electronics (Promotion consumption goal) Cell Phones and accessories (Promotion consumption goal)
Average helpfulness	Health and personal care (Prevention consumption goal) Grocery and gourmet food(Prevention consumption goal) Office product (Promotion consumption goal) Baby (Prevention consumption goal) Beauty (Prevention consumption goal) Pet supplies (Prevention consumption goal)

Table 4.30: Two most important distinguishing features and their respective categories

The value of overall is the most important factor in classification of reviewers for products such as Books, Electronics, and Cell phones and accessories. Average overall is the rating that reviewer themselves assign to the product being reviewed. A reviewer with high average overall is someone who mostly purchases good products and shares satisfactory experiences in the reviews. Whereas a reviewer with lower overall is someone who mostly purchases bad products and shares dissatisfactory experiences in the reviews. This finding supports the conclusion that we made in Section 4.3.2.3 in Table 4.27—*higher overall leads to helpful reviews for products like Books, Electronics, and Cell phones and accessories*. For products like Books, Electronics, and Cell phones and accessories, overall is the most important differentiating factor to determine helpfulness and classify reviewers.

4.4.2 Classification accuracy

In this section, we present and analyze the classification accuracies of the J48 classifiers in each product category. We use a 10-fold cross validation technique to estimate the classification accuracies. In a 10 fold cross validation, the original dataset is randomly partitioned into 10 equal-size sub-datasets. Out of the 10 sub-datasets, 9 sub-datasets are used as training data and the remaining 1 sub-dataset is retained as the validation data for testing the classifier. This process is repeated for 10 times, with each sub-datasets used exactly once as the validation data. The result is then summed to produce a single estimation.

Then we generate a confusion matrix to measure the performance of each classification model. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (Powers, 2011).

For the Books category, for example, we can see that the J48 classifier is highly accurate with 99.82% correct predictions, as shown in Table 4.31. For each class of reviewers—expert, conscientious and novice, the true positive rate is 0.998 or better. This shows that the feature set we chose for clustering is very effective in classifying reviewers into different classes. The confusion matrix for this experiment is presented in Table 4.31 and other class-wise precision analysis like true positive, false positive, precision, recall, and F-measure are presented in Table 4.32.

Classified as	C1 (conscientious)	C2 (expert)	C3 (novice)
C1 (conscientious)	93307	43	88
C2 (expert)	88	128938	138
C3 (novice)	76	128	92517

Table 4.31: Confusion matrix for the "Books" category.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
C1 (conscientious)	0.999	0.001	0.998	0.999	0.998
C2 (expert)	0.998	0.001	0.999	0.998	0.998
C3 (novice)	0.998	0.001	0.998	0.998	0.998

Table 4.32: Detailed accuracy by class for the "Books" category.

The confusion matrices for all the categories can be found in the Appendix, Section D.

The weighted average accuracy for each category is listed in Table 4.33.

Category list	TP Rate	FP Rate	Precision	Recall	F-Measure
Books	0.998	0.001	0.998	0.998	0.998
Electronics	0.998	0.001	0.998	0.998	0.998
Cell Phones and accessories	0.997	0.004	0.997	0.997	0.997
Health and personal care	0.999	0.000	0.999	0.999	0.999
Grocery and gourmet food	0.995	0.002	0.995	0.995	0.995
Office product	0.998	0.001	0.998	0.998	0.998
Baby	0.966	0.012	0.966	0.966	0.966
Beauty	0.998	0.001	0.998	0.998	0.998
Pet supplies	0.997	0.001	0.997	0.997	0.997

Table 4.33: Weighted average accuracies of all categories.

The key for Table 4.33 appears below:

TP Rate – the true positive rate in terms of correctly identifying reviewer class (true positive / (true positive + false positive)).

FP Rate – the false positive rate in terms of incorrectly identifying reviewer class (false positive / (true positive + false positive)).

Precision – the precision is the fraction of positively classified reviewers that are relevant.

Recall – the recall is the fraction of relevant reviewers that are classified.

F-measure – the weighted average of precision and recall, where 1 is the best score and 0 is the worst score.

4.5 Summary

In this Chapter, we performed clustering and classification to demonstrate how we were able to find different classes of reviewers and different attributes in the reviews that affected their perceived helpfulness across various product types. Using the helpfulness as a quality metric and frequency of review as a quantity metric we demonstrated that

reviewers can have highest level of maturity known as *expert reviewers* and lowest level of maturity known as *novice reviewers* and any level between them.

In Section 4.3, we performed clustering on reviewers and then labeled these clusters into different classes that reflect the expertise of reviewers such as novice, conscientious and experts, based on the features of each clusters. We then use this data to build classifier for categorizing reviewers into one of the aforementioned classes. Hence we achieved Objective 1: Demonstrate that reviewers can be either expert or novice by performing data clustering and then doing data analysis to identify attributes that make them expert or novice. We will use the quality and quantity of reviews as metrics to define expert and novice reviewers. Differentiating different classes of reviewers will help to further understand the behavior of each class over time and over different product type.

In Section 4.4, we created a J48 decision tree to identify the rules or features that help to predict reviewer class. We achieved Objective 2: Demonstrate that a number of features like review length, overall (rating), helpfulness, etc. affect review classification by developing decision tree for data classification to find features that differentiates clusters from one another. Understanding what roles which features play more importantly than others to perform classification will help us more accurately predict reviewer class.

In Section 4.3.2.3, we analyzed reviewers across different product categories to understand if the perceived helpfulness of a review is affected by product categories. These products are diverse in terms of their consumed goal and usage. Hence we achieved Objective 3: Demonstrate that reviews are valued differently across different

product categories by performing clustering on reviewers on diverse product categories such as (1) Books, (2) Electronics, (3) Cellphones and accessories, (4) Health and personal care, (5) Grocery and gourmet food, (6) Office product, (7) Baby, (8) Beauty, and (9) Pet supplies. Understanding that reviews are valued differently across different product types will help us identify salient features for each category, implying that any recommendation systems would have to consider different products could demand different solutions.

Utilizing the aforementioned findings, we have obtained insights for building our proposed recommendation system framework that centers around (1) reviewers, and (2) products. The recommendation system framework is capable of training reviewers by recommending actions that would help them write better quality reviews. As stated in Objective 1, reviewers differ from one another in the level of reviewing expertise ,varying from *expert* level to *novice* level. Therefore, *different classes of reviewers may need to work on different skill sets to become better at reviewing*; for example, some reviewers may require to write review frequently whereas others may require to write informative reviews. In Chapter 5, we will first investigate if each class of reviewers evolve over time and then train reviewers who are lagging behind by leveraging the actions of expert reviewers.

As stated in Objective 3, products vary from one another as they are associated with different consumption goals- *promotion* and *prevention*. Reviews targeted for products with different consumption goals need to emphasize on different features; for example, *review length* for some categories whereas *overall* for others. Therefore, in Chapter 5, we will emphasize on review features according to the product category when

we design recommendation system framework. We will be using product related findings that are highlighted in Table 4.34 to design recommendation system framework in Chapter 5.

Findings	References
Higher overall leads to helpful reviews for products associated with promotion consumption goal such as Books, Electronics, Cell Phones and accessories, and Office product	H1 (Table 4.27)
Longer review leads to helpful reviews for products associated with prevention consumption goal such as Grocery and gourmet food, Health and personal care, Baby, Beauty, and Pet supplies.	H2 (Table 4.27)
Expertise for reviewing Grocery and gourmet food does not necessarily come with practice and experience.	Conclusion 4d

Table 4.34: Product related finding and their references.

Chapter 5
Recommendation System Framework

As stated in Chapter 1, one of our main goals in this Thesis after differentiating reviewers based on the *quality* of their reviews is to devise an approach to leverage expert reviewers' behaviors to help train novice reviewers effectively and efficiently. In this chapter we propose a recommendation system framework that trains novice to follow the action sequence of expert in order to improve their reviewing skill.

First, we find whether each class of reviewers evolves or changes over time, in terms of their reviewing skills, and if yes, how each class changes. Understanding reviewer evolution process will provide insights on how reviewers are evolving on their own without any training. Also, we can answer which phase of their evolution our recommendation system framework should facilitate. Second, we perform a sentiment analysis on the review text to understand the tone used by different classes of reviewers. We want to understand whether different classes of reviewers use a different tone and how that affects review helpfulness. McAuley *et al*., (2013) points out that there is a strong relation with "expertise" from the light of linguistic development. If there is a relation between review tone and review helpfulness in expert reviewer class, then we can learn from experts and make appropriate tone recommendations to novices and help them write better reviews. Third, we present an architecture that recommends actions

from experts to train the reviewer who is lagging behind. We pursue the following series of objectives below:

1. **Objective 4:** Demonstrate that reviewers evolve with time by performing graph analysis of review helpfulness over time. The pattern of reviewer evolution would present insights on how different classes of reviewers evolve with time. It would help to answer if it is possible for a novice or conscientious reviewer to become an expert reviewer with time.
2. **Objective 5:** Demonstrate that expert, novice, and conscientious reviewers use different tones while reviewing by performing sentiment analysis on review text. The sentiment analysis of review texts would present insights on which tone is used by experts in certain product type. Novice reviewers can be recommended to use the same tones as the experts to make their reviews more helpful.
3. **Objective 6:** Demonstrate that actions of experts can be leveraged to train novice reviewers to write good quality reviews. Experts' actions together with the product-specific review features are combined to make appropriate review recommendations.

5.1 User evolution

We know that helpfulness is a measure of how useful/helpful the review is from other users' perception. Helpfulness of a review measures the worthiness of the review to other users. So we investigate the trend of helpfulness in different classes to find if reviewers are maturing—i.e., becoming more expert in writing useful/helpful reviews—over time.

Our objective in this experiment is Objective 4: Demonstrate that reviewers evolve with time by performing graph analysis of review helpfulness over time. The

pattern of reviewer evolution would present insights on how different classes of reviewers evolve with time. It would help to answer if it is possible for a novice or conscientious reviewer to become an expert reviewer with time.

5.1.1 Setup

We first detect outliers, and remove them from dataset, before plotting average helpfulness of reviewers over time. Outliers may cause a negative effect on data analyses, or may provide useful information about data when we look into an unusual response to a given study (Seo, 2006). We want to observe the evolution trend of each class–expert, novice and conscientious, thus it is important that we remove outliers of each class to understand the evolution trend of core reviewers. In each class, majority of reviewers follow similar behavior in terms of frequency of reviewing, review length, review helpfulness, review rating and active month, but there might be a small number of outliers that have unusually large or small values compared to others in the same class or cluster.

First, we detect outliers by determining an interval spanning over the mean plus/minus two standard deviations. 95.45% of values lie within a band around the mean in normal distribution with a width of two standard deviation. Below is the mathematical notation, where x is any observation from normally distributed random values, μ is the mean of the distribution and σ is its standard deviation.

$$pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545$$

We detect and remove outliers from each class of all categories. Table 5.1 shows reviewer count in each class before and after removing outliers.

Category	Reviewer class	Reviewer count before removing outliers	Reviewer count after removing outliers	Percentage of reviewer count after removing outliers
Books	C1 (conscientious)	419721	333757	79.518%
	C2 (expert)	559123	462749	82.763%
	C3 (novice)	353758	295907	83.646%
Electronics	C1 (novice)	106666	86604	81.191%
	C2 (expert)	205839	170339	82.753%
	C3 (conscientious)	231271	197256	85.292%
Cell Phones & accessories	C1(conscientious)	39018	20786	81.682%
	C2(expert)	26050	20919	80.303%
Health & personal care	C1 (frequent positive expert)	62264	51037	81.968%
	C2 (non-frequent negative expert)	29855	23514	78.760%
	C3 (conscientious)	61970	54223	87.498%
	C4 (novice)	39792	32145	80.782%
Grocery & gourmet food	C1 (conscientious)	35326	30663	86.800%
	C2 (positive expert)	21384	17036	79.667%
	C3 (novice)	29285	23415	79.955%
	C4 (negative expert)	15101	12107	80.173%
Office product	C1 (expert)	22032	17691	80.296%
	C2 (novice)	10047	7675	76.390%
	C3 (conscientious)	35377	28919	81.745%
Baby	C1 (novice)	3152	3152	100%
	C2 (frequent positive expert)	5573	5573	100%
	C3 (conscientious)	5063	5063	100%
	C4 (non-frequent negative expert)	3055	3055	100%
Beauty	C1 (positive expert)	12138	10607	87.386%
	C2 (negative expert)	6728	5479	81.435%
	C3 (conscientious)	14803	13394	90.481%
	C4 (novice)	9116	7646	83.874%
Pet supplies	C1 (frequent positive expert)	7730	6585	85.187%
	C2 (non-frequent negative expert)	4040	3296	81.584%
	C3 (conscientious)	11393	10218	89.686%
	C4 (novice)	6553	5499	83.915%

Table 5.1: Reviewer count in each class before and after removing outliers

Then, after removing the outliers, we find the average helpfulness value of each year for each class. This allows to plot a graph of how average helpfulness changes over time for each year. If helpfulness in a class increases then we can say that reviewers in the class mature over time. Whereas if helpfulness of a class decreases then we can say that reviewers in the cluster do not mature over time. We perform this analysis for all nine categories.

5.1.2 Discussion

Here we report on our change analysis of helpfulness for all nine product categories which are divided into two sets of products—promotion consumption goal and prevention consumption goal—based on Section 4.3.2.3.

We first perform this analysis on products associated with *promotion consumption goal* such as Books, Electronics, Cellphones and accessories, and Office product. Here we present the results on Books, and provide the others in the Appendix, Section E. From Chapter 4, we know that there are three classes of reviewers in Books: (1) conscientious, (2) expert, and (3) novice. For each class, we calculate the average helpfulness of all reviews written by reviewers belonging to the particular class with respect to time (in year). Table 5.1 shows the average helpfulness values of the three classes for each year.

	Average helpfulness		
Year count	**Conscientious**	**Expert**	**Novice**
1	0.744	0.877	0.787
2	0.754	0.891	0.806
3	0.758	0.897	0.817
4	0.759	0.898	0.820
5	0.762	0.896	0.823
6	0.765	0.894	0.821
7	0.768	0.894	0.812
8	0.768	0.894	0.816
9	0.787	0.891	0.811
10	0.772	0.893	0.828
11	0.786	0.900	0.814
12	0.777	0.935	0.836
13	0.801	0.885	0.825
14	0.854	0.911	0.856
15	0.884	0.935	0.845
16	0.979	0.959	
17		0.967	

Table 5.2: Average helpfulness of three clusters in each active year for "Books" category

We can observe from Table 5.2 and its respective plot in Figure 5.1 that the average helpfulness of all three classes of reviewers increases with time but the rate of this growth

is different for different classes. Below are the linear equations for each trend line and error in Books.

$$Conscientious: y = 0.010x + 0.709;\ R^2 = 0.685$$

$$Expert: y = 0.003x + 0.87;\ R^2 = 0.596$$

$$Novice: y = 0.002x + 0.799;\ R^2 = 0.597$$

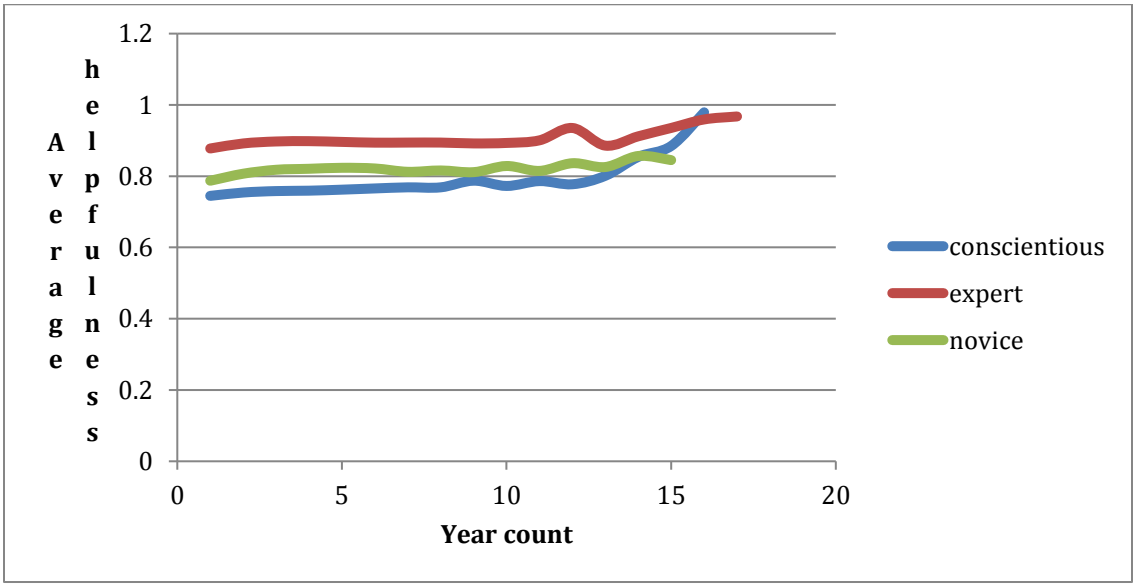

Figure 5.1: Trend of average helpfulness over time for three clusters in "Books" category

	Conscientious Vs. Expert	Conscientious Vs. Novice	Expert Vs. Novice
p- value	**4.2756E-08**	**0.0221**	**3.5787E-13**

Table 5.3: *t*-Test result for change on helpfulness in clusters in "Books" category. Bolded values (all) are statistically significant, $p < 0.05$.

The average helpfulness of conscientious reviewer increases at a faster rate than that of expert reviewers and that of expert reviewers increases faster than novice's. This increase is statistically significant as observed from *t*-test in Table 5.3.

The trend of each classes in products categories related with promotion consumption goal are listed below in Table 5.4. In the beginning, conscientious reviewers have the least average helpfulness compared to other classes. Then they show a gradual increase in average helpfulness over time. After 15 years, they reach the same

level as the expert reviewers. Meanwhile, novices generally start with a higher average helpfulness value than conscientious but they never grow enough to reach the same level as expert reviewers even after 15 years of reviewing.

Taking together Tables 5.3 and 5.4, we see that the *p*-value is less than 0.05 which verifies that helpfulness of conscientious reviewer grows *faster* than expert reviewer ($0.010 > 0.003$, Table 5.4). Similarly, helpfulness of expert reviewers grows *faster* than novices ($0.003 > 0.002$, Table 5.4).

Category	Reviewer class	Linear equation of trend line	R^2 error
Books	C1 (conscientious)	$y = 0.010x + 0.709$	0.685
	C2 (expert)	$y = 0.003x + 0.872$	0.596
	C3 (novice)	$y = 0.002x + 0.799$	0.597
Electronics	C3 (conscientious)	$y = 0.007x + 0.870$	0.690
	C2 (expert)	$y = 0.0016x + 0.909$	0.837
	C3 (novice)	$y = 0.0015x + 0.861$	0.683
Cell Phones & accessories	C1 (conscientious)	$y = 0.014 + 0.802$	0.728
	C2 (expert)	$y = 0.007x + 0.868$	0.778
Office product	C3 (conscientious)	$y = 0.015x + 0.834$	0.706
	C1 (expert)	$y = 0.0016x + 0.912$	0.711
	C2 (novice)	$y = 0.0015x + 0.853$	0.786

Table 5.4: Linear equation of trend line for each reviewer class in product categories belonging to promotion consumption goal.

In Table 5.4, fourth column denotes *coefficient of determination* referred as R^2 error which is a statistical measure of how well observed outcomes i.e., y are predicted by the model based on the variance in the outcomes explained by the model (Draper and Smith, 1998). In other words, how close the data are to the fitted trend line. The value of R^2 error ranges from 0 to 1 where 0 denotes that the dependent variable (y) cannot be predicted from the independent variable (x), and 1 denotes that the dependent variable (y) can be predicted without error from the independent variable (x). In Table 5.4, we can observe that trend line predicts least accurately for Books experts (59%) and most accurately for Electronics experts of (83%). For all reviewer classes R^2 error is greater

than or equal to 0.59 which means that 59 percent or more of the variation in y is predictable from x. Looking at R^2 error we can tell that although the predicted y in Table 5.4 are not 100% accurate, it is a good prediction because the predicted y is the average helpfulness which is the number of "up" votes provided by human readers who find the review helpful. Because it is very difficult to predict human actions the R^2 error of greater than or equal to 0.59 is a very good prediction. In general, for products associated with promotion consumption goal from Table 5.4, we observe that:

- Increase in average helpfulness of conscientious reviewers is the fastest compared to other classes.
- Increase in average helpfulness of expert reviewers is slower than conscientious reviewers but faster than novice reviewers.
- Average helpfulness of novice remains constant.

We performed *t*-test to check the statistical significance of the above observations. The details of this below in Table 5.5:

Category	Conscientious Vs. Expert	Conscientious Vs. Novice	Expert Vs. Novice
Books	**4.2756E-08**	**0.0221**	**3.5787E-13**
Electronics	0.3580	**0.0001**	**2.3980E-07**
Cell Phones & accessories	0.1631	NA (only 2 clusters)	NA (only 2 clusters)
Office product	0.0717	**0.0078**	**0.0017**

Table 5.5: *t*-Test result for change in helpfulness in each reviewer class in product categories belonging to promotion consumption goal. Bolded values are statistically significant, $p < 0.05$.

Note that in Cell phones and accessories, there are only two clusters – conscientious and experts. From Table 5.5, we can see that the increase in average helpfulness of conscientious cluster is not statistically significant when compared with expert cluster. So based on *t*-test and the trend lines in Table 5.4 we can make the following conclusions:

- Conscientious: Increase in average helpfulness of conscientious reviewers is either *faster* than (Books: $0.010 > 0.003$, Table 5.4) or *similar* to expert reviewers (Electronics: $0.007 \cong 0.001$; Cell Phones and accessories: $0.014 \cong 0.007$; Office product: $0.015 \cong 0.001$, Table 5.4).
- Expert: Increase in average helpfulness of expert reviewers is always *faster* than novice users (Books: $0.003 > 0.002$; Electronics: $0.0016 > 0.0015$; Office product: $0.0016 > 0.0015$, Table 5.4).
- Novice: Increase in average helpfulness of novice reviewers is the *least* with respect to conscientious and expert cluster (Books: 0.002; Electronics: 0.0015; Office product: 0.0015, Table 5.4).

Having performed the analysis on promotion consumption goal-related products, we now perform the same analysis on products related to prevention consumption goal such as Health and personal care, Grocery and gourmet food, and Baby product. Here we look at Health and personal care products to illustrate our findings. Details about other products can be found in Appendix, Section E.

Now, we know that, from Table 4.7, there are four classes of reviewers in Health and personal care: (1) novice, (2) conscientious, (3) frequent positive expert, and (4) non-frequent negative experts.

For each class, we generate Figure 5.2 in the same way as we did with Figure 5.1.

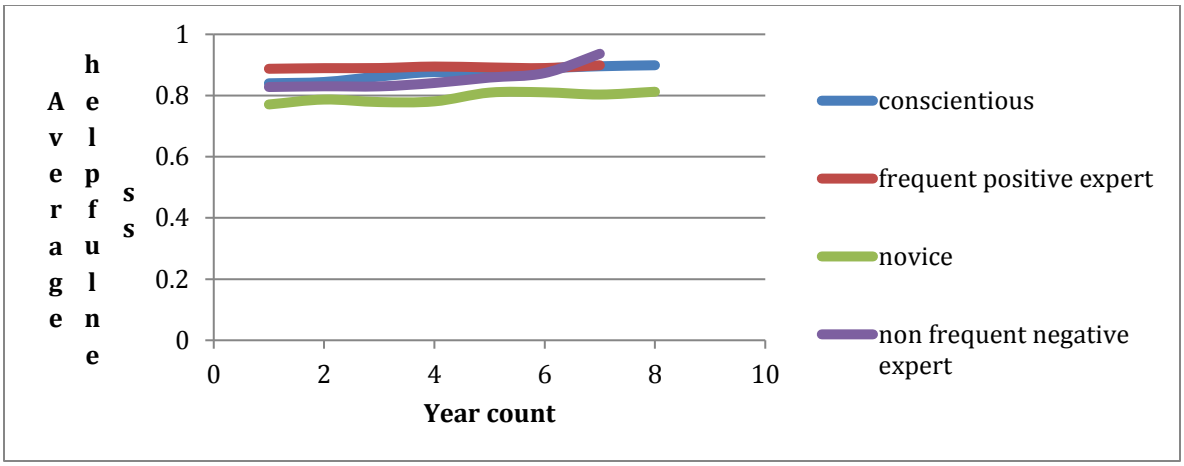

Figure 5.2: Trend of average helpfulness over time for three clusters in "Health and personal care" category

Below are the linear equations of trend line and error from Figure 5.2.

$$Conscientious: y = 0.012x + 0.832;\ R^2 = 0.926$$

$$Frequent\ positive\ expert: y = 0.008x + 0.886;\ R^2 = 0.641$$

$$Novice: y = 0.005x + 0.767;\ R^2 = 0.760$$

$$Non - frequent\ negative\ expert: y = 0.015x + 0.793;\ R^2 = 0.759$$

The increase in average helpfulness is the fastest in non-frequent negative expert reviewers, followed by conscientious reviewers, followed by frequent positive expert reviewers, and rounded out by novice reviewers. At the start, non-frequent negative expert reviewers have average helpfulness less than both frequent positive expert reviewers and conscientious reviewers. Then, they show a gradual increase in average helpfulness over time. After 7 years, they surpass frequent positive expert and conscientious reviewers. Meanwhile, novices start with the least average helpfulness and they remain as such for 8 years. We then perform *t*-test to check if the observations we made are statistically significant. Details of *t*-test are presented in Table 5.6:

	Conscientious vs. Frequent positive expert	Conscientious vs. Non-frequent negative expert	Conscientious vs. Novice	Frequent positive expert vs. Non-frequent negative expert	Non-frequent negative expert vs. Novice	Frequent positive expert vs. Novice
p-value	**0.0074**	0.3323	**8.6786E-07**	**0.0268**	**0.0003**	**1.441E-07**

Table 5.6: *t*-Test result for change in helpfulness in clusters in "Health and personal care" category. Bolded values are statistically significant, $p < 0.05$.

From Table 5.6 and 5.7, we can see that increase in average helpfulness of conscientious and non-frequent negative expert are *not* statistically significant i.e., they both grow at a *similar* rate ($0.012 \cong 0.015$, Table 5.7). On the other hand, the other differences are statistically significant. Frequent positive expert reviewers grow at a *slower* rate than both conscientious and non frequent negative expert reviewers ($0.008 < 0.012; 0.008 < 0.015$, Table 5.7); and at a *faster* rate than novice reviewers ($0.008 > 0.005$, Table 5.7). Novice reviewers have the *slowest* growing rate (0.005, Table 5.7).

Finally, the trends and errors of all classes in product categories related with prevention consumption goal are listed below in Table 5.7.

Category	**Reviewer class**	**Linear equation of trend line**	R^2 **error**
Grocery & gourmet food	C1 (conscientious)	$y = 0.018x + 0.846$	0.920
	C2 (positive expert)	$y = 0.010x + 0.833$	0.744
	C3 (novice)	$y = 0.001x + 0.889$	0.707
	C4 (negative expert)	$y = 0.013x + 0.766$	0.742
Health & personal care	C3 (conscientious)	$y = 0.012x + 0.832$	0.926
	C1 (frequent positive expert)	$y = 0.008x + 0.886$	0.641
	C4 (novice)	$y = 0.005x + 0.767$	0.760
	C2 (non-frequent negative expert)	$y = 0.015x + 0.793$	0.759
Baby	C3 (conscientious)	$y = 0.012x + 0.782$	0.555
	C2 (frequent positive expert)	$y = -0.008x + 0.862$	0.416
	C1 (novice)	$y = -0.010x + 0.812$	0.196
	C4 (non-frequent negative expert)	$y = -0.007x + 0.8900$	0.241
Pet supplies	C1 (frequent positive expert)	$y = 0.0098x + 0.912$	0.740
	C2 (non-frequent negative expert)	$y = 0.0097x + 0.8642$	0.865
	C3 (conscientious)	$y = 0.013x + 0.8887$	0.661
	C4 (novice)	$y = 0.0091x + 0.8371$	0.623
Beauty	C1 (positive expert)	$y = 0.0127x + 0.861$	0.840

	C2 (negative expert)	$y = 0.0206x + 0.7902$	0.586
	C3 (conscientious)	$y = 0.058x + 0.8516$	0.809
	C4 (novice)	$y = 0.0017x + 0.796$	0.583

Table 5.7: Linear equation of trend line for each reviewer class in product categories belonging to prevention consumption goal.

From Table 5.7, we observe that like Health and personal care, average helpfulness of conscientious cluster grows fastest in both Grocery and gourmet food and Baby product, followed by negative experts. In Table 5.7, we can observe from R^2 error that trend line predicts least accurately for Baby novice reviewers (0.19) and most accurately for Grocery and gourmet food conscientious reviewers of (0.92). However on careful speculation we can see that R^2 error for all clusters except Baby is greater than 0.64 which is very good prediction. Details of our *t*-test results are presented in Table 5.8.

Category	**Conscientious vs. Positive Expert**	**Conscientious vs. Negative Expert**	**Conscientious vs. Novice**	**Positive vs. Negative**	**Negative vs. Novice**	**Positive Expert vs. Novice**
Grocery & gourmet food	**0.0005**	**1.7346E-05**	**0.0123**	0.113	**0.0102**	**0.0167**
Beauty	**0.0034**	**0.4781**	**1.3352E-07**	0.0721	**0.0044**	**3.7337E-07**
Category	**Conscientious vs. Frequent positive expert**	**Conscientious vs. Non-frequent negative expert**	**Conscientious vs. Novice**	**Frequent positive expert vs. Non-frequent negative expert**	**Non-frequent negative expert vs. Novice**	**Frequent positive expert vs. Novice**
Health & personal care	**0.0074**	0.3323	**8.6786E-07**	**0.0268**	**0.0003**	**1.441E-07**
Pet supplies	**0.0035**	**0.0159**	**0.0005**	**0.0011**	**0.0176**	**2.7577E-05**
Baby	0.3448	0.1185	**0.0337**	**0.0020**	**2.6023E-05**	**0.0022**

Table 5.8: *t*-Test result for change in helpfulness in clusters in product categories belonging to prevention consumption goal. Bolded values are statistically significant, $p < 0.05$.

For products related with prevention consumption goal, we can make the following conclusions from Table 5.7 and *t*-test results in Table 5.8 which is true for all the product categories:

- Conscientious: Increase in average helpfulness of conscientious expert reviewers is always *faster* than novice (Grocery and gourmet food: $0.018 > 0.001$; Health and personal care: $0.012 > 0.005$; Baby: $0.012 > -0.010$, Beauty: $0.058 > 0.001$, Pet supplies: $0.013 > 0.009$, Table 5.7).
- Non-frequent negative expert: Increase in average helpfulness of non-frequent negative expert reviewers is also always *faster* than novice (Grocery and gourmet food: $0.013 > 0.001$; Health and personal care: $0.015 > 0.005$; Baby: $-0.007 > -0.010$, Beauty: $0.020 > 0.001$, Pet supplies: $0.0097 > 0.0091$, Table 5.7).
- Frequent positive expert: Increase in average helpfulness of frequent positive expert reviewers is again also always *faster* than novice (Grocery and gourmet food: $0.010 > 0.001$; Health and personal care: $0.008 > 0.005$; Baby: $-0.008 > -0.010$; Beauty: $0.012 > 0.001$; Pet supplies: $0.0098 > 0.0091$, Table 5.7).

Some other product-specific observations that we can make from Table 5.7 and *t*-test results in Table 5.8 are:

- Conscientious: For Health and personal care, increase in average helpfulness of conscientious is *similar* to non-frequent negative expert ($0.012 \cong 0.015$, Table 5.7); and at a *faster* rate than frequent positive experts ($0.012 > 0.008$, Table 5.7). For Baby, increase in average helpfulness of conscientious is *similar* to both non-frequent negative expert ($0.012 \cong -0.007$, Table 5.7) and frequent positive experts ($0.012 \cong -0.008$, Table 5.7). However for Grocery and gourmet food, increase in average helpfulness of conscientious is *faster* than both negative expert ($0.018 > 0.013$, Table 5.7) and positive expert ($0.018 > 0.010$, Table 5.7). For Beauty products, increase in average helpfulness of conscientious is

similar to negative expert ($0.058 \cong 0.0206$, Table 5.7) and *faster* than positive experts ($0.058 > 0.0127$, Table 5.7). For Pet supplies, increase in average helpfulness of conscientious is *faster* than both non-frequent negative expert ($0.013 > 0.0097$, Table 5.7) and frequent positive experts ($0.012 > 0.0098$, Table 5.7).

- Non-frequent negative expert: For Health and personal care, and Baby, increase in average helpfulness of non-frequent negative expert reviewers is always *faster* than positive frequent experts (Health and personal care: $0.015 > 0.008$; Baby: $-0.007 > -0.008$, Table 5.7). For Grocery and gourmet food, and Beauty products, increase in average helpfulness of negative expert reviewers is *similar* to the positive expert reviewers (Grocery and gourmet food: $0.013 \cong 0.010$; Beauty: $0.020 \cong 0.012$, Table 5.7). For pet supplies, increase in average helpfulness of non-frequent negative expert reviewers is always *slower* than positive frequent experts (Pet supplies: $0.0097 < 0.0098$, Table 5.7). Hence, average helpfulness of non-frequent negative expert reviewers may be *faster* than , *similar* to, or slower than positive frequent experts for products related to prevention consumption goal.

From above observations, for products related to both promotion and prevention consumption goal we can make two important conclusions:

Conclusion 5a: Average helpfulness of the conscientious and expert clusters (including both the frequent positive and non-frequent negative) grow faster than that of the novice cluster.

Conclusion 5b: Average helpfulness of the conscientious cluster increases faster than or similar to that of the expert cluster.

Also, expert clusters (i.e., frequent positive experts and non-frequent negative experts) may grow at either *similar* rate or *faster* than each other. Since experts clusters in products related with prevention consumption goal follow different trends, we cannot claim a strong relation on how frequent positive experts and non-frequent negative experts grow relative to each another. *The relationship between positive experts and negative experts are category-specific for prevention consumption goal-related products*.

Note that for Baby products, average helpfulness of all classes of reviewers except conscientious cluster decrease over time. This is unique to Baby products and it indicates ***that reviewers are posting worse reviews with time.*** One possible explanation could be due to a very low reviewer count in Baby (6033, Table 4.21 in Chapter 4) when compared with other prevention consumption goal related products (e.g., Health and personal care = 70669, Table 4.15 in Chapter 4; Grocery and gourmet food = 34711, Table 4.17 in Chapter 4; Beauty = 42785, Table 4.23 in Chapter 4; Pet supplies = 29716, Table 4.25 in Chapter 4). Further investigations are needed to find factors that could have caused the decrease of review quality (i.e., average helpfulness) overtime in Baby products. With this in mind, for future reference, we should be cautious to perform experiment on only product categories with high unique reviewer count. Also, from Table 5.7 we can see that R^2 error for reviewer clusters except conscientious, in Baby are comparatively small (less than 0.42) with respect to all other clusters from product categories related to both promotion and prevention consumption goal. Hence these trend lines are the least accurate compared to the other trend lines that we have obtained for all

the other products. For these reasons, we will exclude observations of insights of the Baby products from our framework.

5.1.3 Result

With time, users consume more products and their tastes change or in other words they become experienced (McAuley, 2013). This is true for all the classes of reviewers. Irrespective of product types, we have observed that *reviewers evolve with time*. *Conscientious reviewers, as the name implies, are known for their diligence for reviewing*. As stated by conclusion 5b, the growth rate of conscientious in some product categories such as Books, Grocery and gourmet food, and Pet supplies is the highest compared to all other clusters. However in other categories, the growth rate of conscientious is *similar* to that of experts. *For conscientious reviewers, their interest and diligence could reason for their quicker learning ability to write helpful reviews compared to other classes*.

From conclusions 5a and 5b, we know how each classes of reviewers evolve, and at what rate do they evolve relative to one another. We can use this information to learn which action sequence in terms of review features and reviewing frequency leads to what kind of evolution. For example, a cluster who posts review more frequently or more elaborately (or lengthily) may grow faster than the others. We can then *recommend the actions of the cluster which grows quickly to the cluster which grows slowly*. For example, the actions of experts can be recommended to conscientious and the actions of conscientious can be recommended to novices. Based on the findings of user evolution, we propose a recommendation framework in detail in Section 5.3.

There are other factors that affect the evolution. For example, the evolution in terms of ratings behavior of a reviewer may be the results of number of factors such as shifting trends in community, arrival of new products, and even change in the users' social network (McAuley, 2013). Another explanation could be related to linguistic difference in review text between different classes of reviewers—novice, expert and conscientious. There is a strong relation with "expertise" from the light of linguistic development (Romaine, 1984). We have to consider the effect of language used by different classes of reviewers. Do experts use more pronouns than novice? Are conscientious reviews more personal by using "I" instead of "We"? Do conscientious reviewers have positive tone than expert reviewers? To answer these questions and further understand how the tone preference changes in each cluster for different product types, we perform sentiment analysis in review text.

5.2 Sentiment analysis

We know that reviews present reviewers' opinion based on the experience with the consumed product. The opinions may be positive, negative or neutral. Sentiment Analysis labels people's opinions as different categories such as positive and negative from a given piece of text (Madhoushi *et al*., 2015).

Our objective in this experiment is Objective 5: Demonstrate that expert, novice, and conscientious reviewers use different tones while reviewing by performing sentiment analysis on review text. The sentiment analysis of review texts would present insights on which tone is used by experts in certain product type. Novice reviewers can be recommended to use the same tones as the experts to make their reviews more helpful.

5.2.1 Setup

We use VADER to find the *polarity* and *intensity* of all product reviews. As covered in Chapter 2, sentiment *polarity* may be positive, negative, or neutral and *intensity* may range from -4 to +4. For each reviewer, the average of the intensities in each polarity is computed for all the reviews posted by the reviewer. Therefore, for each reviewer their opinions are measured in an average positive intensity, average negative intensity, average neutral intensity, average compound intensity. We then aggregate the sentiment of all reviewers in each class to compute the class average sentiment. Then the next step is to look into the relation between review polarity and other features such as review length, average helpfulness, and average overall to understand their interdependencies, if any.

5.2.2 Discussion

For each class of reviewers, we calculate correlation between review features such as average review length, average helpfulness, and average overall with review sentiments measured by average positive intensity, average negative intensity, average neutral intensity, and average compound intensity. Below are some of the observations:

- Average helpfulness and review sentiment: There is a very low correlation of average helpfulness with (1) positive tone (-0.096), (2) negative tone (0.031), and (3) neutral tone (0.08). The product-specific details are in Appendix, Section F. This suggests that *there is no significant relation between review sentiment and average helpfulness*.
- Average overall and review sentiment: There is a high correlation between average overall and review sentiment. Specifically, the correlation between

average overall with (1) positive tone is 0.297 and (2) negative tone is -0.282. The product-specific correlation details are in Appendix, Section F. This signifies that a satisfied consumer (i.e., high average overall) shares more positive experiences (i.e., high positive tone) with the product in their reviews. Conversely, a dissatisfied consumer provides low rating to products (i.e., high average overall) and has more negative tone in their reviews. Hence this is a verification that *a satisfied consumer who provides high overall has more positive information in their reviews which is reflected by the positive tone*.

- Review length and review sentiment: *The correlation between review length and review sentiment is more pronounced* and we can make number of observations from this relation. Below Table 5.9 displays the correlation between review length and positive, negative, and neutral tone within each classes of reviewers.

		Review length		
Product Category	**Clusters**	**Positive**	**Negative**	**Neutral**
Books	C1 (conscientious)	-0.282	0.204	0.159
	C2 (expert)	-0.375	0.308	0.234
	C3 (novice)	-0.325	0.285	0.205
Cell Phones & accessories	C1 (conscientious)	-0.351	0.080	0.343
	C2 (expert)	-0.249	0.008	0.257
Electronics	C1 (novice)	-0.128	-0.013	0.125
	C2 (expert)	-0.256	0.099	0.214
	C3 (conscientious)	-0.288	0.104	0.251
Office product	C1 (expert)	-0.302	0.096	0.263
	C2 (novice)	-0.065	-0.046	0.091
	C3 (conscientious)	-0.269	0.087	0.241
Grocery & gourmet food	C1 (conscientious)	-0.307	0.144	0.262
	C2 (novice)	-0.207	-0.059	0.237
	C3 (positive expert)	-0.346	0.120	0.307
	C4 (negative expert)	-0.179	-0.036	0.194
Health & Personal care	C1 (frequent positive expert)	-0.256	0.104	0.205
	C2 (non frequent negative expert)	-0.129	-0.021	0.134
	C3 (conscientious)	-0.270	0.132	0.212
	C4 (novice)	-0.141	0.008	0.129
Baby	C1 (novice)	-0.200	-0.026	0.211
	C2 (frequent positive expert)	-0.363	0.110	0.334
	C3 (conscientious)	-0.333	0.098	0.317

	C4 (non frequent negative expert)	-0.228	0.082	0.203
Beauty	C1 (positive expert)	-0.284	-0.070	0.253
	C2 (negative expert)	-0.149	-0.092	0.187
	C3 (conscientious)	-0.288	-0.093	0.272
	C4 (novice)	-0.169	-0.098	0.213
Pet supplies	C1 (frequent positive expert)	-0.285	-0.092	0.243
	C2 (non-frequent negative expert)	-0.121	-0.095	0.162
	C3 (conscientious)	-0.305	-0.097	0.277
	C4 (novice)	-0.194	-0.097	0.209

Table 5.9: Correlation between review length and positive, negative, and neutral tone for each cluster.

From Table 5.9 we see that for all clusters in all product categories, correlation of review length with positive tone is higher than neutral tone which in turn is higher than negative tone. We can make following observations from Table 5.9:

- Review length is negatively correlated with positive tone, we can say that *longer reviews are less positive*, in other words, shorter reviews are more positive.
- Review length is positively correlated with neutral tone, we can say that *longer reviews are more neutral*.

5.2.3 Result

The observations we made about the longer reviews being less positive reveals a unique feature of online reviewing process. Usually merchants trying to sell products tend to convey the positive effects of the product. They choose positive adjectives for their online advertising. Hence there is a lot of positive information already in the online advertisement and reviewers do not want to re-iterate the same in their reviews. Rather they would just validate the positive effects of the product as claimed by the advertisement. Hence positive reviews are more likely to be straight and succinct. Hence *shorter reviews are more positive*. Conversely, if a reviewer wants to share negative experience with the product, they have to explain and elaborate their experience as they are pointing the negative product features that are not shared in the product's advertisement. Therefore, *longer reviews are less positive*.

Sentiment analysis provided some insights on the relation between review tone and review features like average helpfulness, average overall, and review length. High correlation of review overall with review sentiment proves that, the tone for all reviewer cluster is solely dependent on the reviewers' experience with the product. Also, the reviewing tone is not different for different class of reviewers. Hence we can conclude that review tone does not have a direct relation with reviewer expertise level.

In this experiment, the tone of each reviewer class such as positive , negative, and neutral is calculated by averaging the tone of all the reviews posted by the reviewers in the class. As a future work, we may want to look at the temporal distribution of review tone for each reviewer class to find if different classes of reviewers use different tone overtime. In other words, we may investigate if there is any change in tone overtime for each reviewer class which may lead to their evolution in terms of review quality (covered in Section 5.2). This might provide insights on the relation between reviewer expertise and review sentiment overtime.

5.3 Recommendation System framework

In this section, we propose a recommendation system framework that provides recommendations in order to help reviewers to improve their reviewing skills and write better quality reviews. The framework works by recommending reviews posted by experts in the past to the reviewer with lesser reviewing expertise.

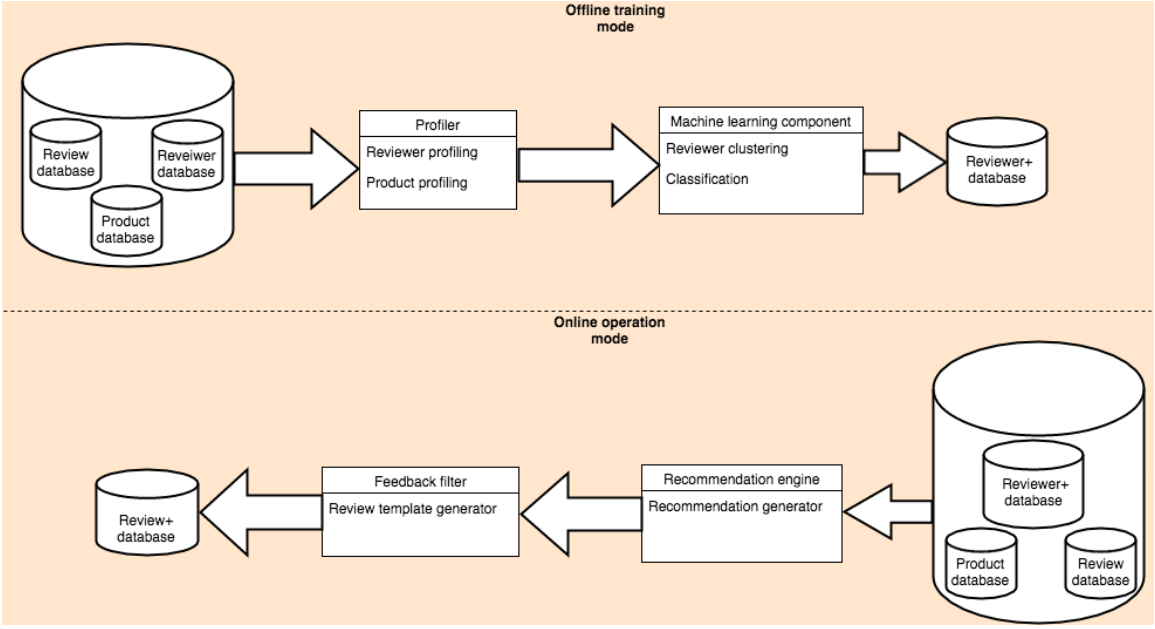

Figure 5.3: Components of Recommendation System framework

Figure 5.3 is the system diagram of recommendation system that we propose. It operates in two modes: (1) offline training mode and (2) online operation mode. It uses multiple databases:(1) a review database, (2) a reviewer database, and (3) a product database. These three databases: review, reviewer, and product databases contain raw data, whereas the ones indicated with a '+' sign are *updated* database with added derived data. For example, the **reviewer+ database** contains the original reviewer database plus *their class label obtained from reviewer clustering* (see Section 4.3 in Chapter 4). In addition, the **review+ database** contains the original review database plus *the usage history of the reviews that have been recommended in the past such as their recommendation counts*. Depending on whether the recommended review helped in increasing the helpfulness of reviews they can be differentiated into two types: (1) reviews that have a record of being helpful in the past, referred to *successful recommendations*, and (2) reviews that have a record of *not* being helpful in the past, referred to *unsuccessful recommendations*. Thus,

the review+ database tags the reviews as *successful recommendations* for future use and *unsuccessful recommendations* that discouraged from future use.

As shown in Figure 5.3, the recommendation system framework is divided into two modes: **offline training** and **online operation**.

As implied by the name, the training mode occurs offline. In training mode, there are two key components:

1. Profiler: This contains reviewer profiler and product profiler. Reviewer profiler extracts feature related to a reviewer such as a total number of reviews posted, total active month, average helpfulness and so on (see Section 4.3.1 in Chapter 4). Product profiler segregates products into promotion and prevention consumption types.
2. Machine Learning Module: This contains clustering of reviewers followed by classification (see Section 4.4 in Chapter 4). The output of this component is a database, referred to as the **reviewer+ database** that contains details of reviewers along with their labels based on the quality of their reviews.

As implied by the name, the operation mode occurs online in real time. In operation mode there are also two key components:

1. Recommendation Engine: This component contains a recommendation generator that uses the **reviewer+ database** to classify a reviewer for whom recommendations are to be generated into one of the classes. For product classification, it uses the product database. Then it computes evolution (average helpfulness vs. active year) of the reviewer and all experts by using the review database. Based on the evolution of the reviewer, it finds the closest experts for

the reviewer. It extracts the reviews posted by the closest experts and then uses product specific feature properties to choose reviews for recommendations. This is covered in detail in Section 5.3.1.1.

2. Feedback Filter: This component contains a review template generator. This review template generator keeps track of *successful recommendations* (recommended reviews which were implemented by reviewers and successfully increased the helpfulness). Additionally, it also flags the recommended reviews, which didn't increase the average helpfulness. At the end of this phase, the **review+ database** is created which has a list of *successful recommendations* as templates that were helpful and will be used in future. The review template database also contains the list of *unsuccessful recommendations* that are flagged from future use because they were not helpful in the past. This is covered in detail in Section 5.3.1.2.

5.3.1 Online operation mode

In the online operation mode, the first step is to generate recommendation followed by feedback process as indicated in Figure 5.3. **Recommendation engine** generates recommendations for a reviewer. The process starts by classifying the reviewer into his or her corresponding class; for example, a reviewer could be an expert, or conscientious, or a novice. Once the class of the reviewer is known, the next step is to **find a group of closest experts** to this reviewer to extract the reviews posted by them in the past. We then classify the product to be reviewed into either the prevention or promotion goal type and use this information to further **extract appropriate reviews** from this group of expert reviewers' reviews. For example, if the product is related to the promotion

consumption goal, then we extract reviews with high overall (rating) as recommendations. This is based on our findings (see Table 4.32 in Chapter 4) that reviews with high overall (rating) have a higher probability of being helpful. The flow chart below illustrates the process of generating recommendations in detail.

Feedback filter generated feedback based on the usage of recommended reviews. Once the recommendations are generated and used by reviewers, the success of the recommendations is measured and is stored for future reference. The recommended reviews that increased the number of up-votes or helpfulness are tagged as *successful recommendations* whereas those recommendations that did not increase the helpfulness of the review are deemed as *unsuccessful recommendations*. The feedback process tracks both *successful* and *unsuccessful recommendations* and stores successful recommendations as templates in the review+ database—i.e., the original review database plus the usage history—for future recommendations. The feedback process is a filtering mechanism in which *successful recommendations* are re-used and *unsuccessful recommendations* are flagged from future use.

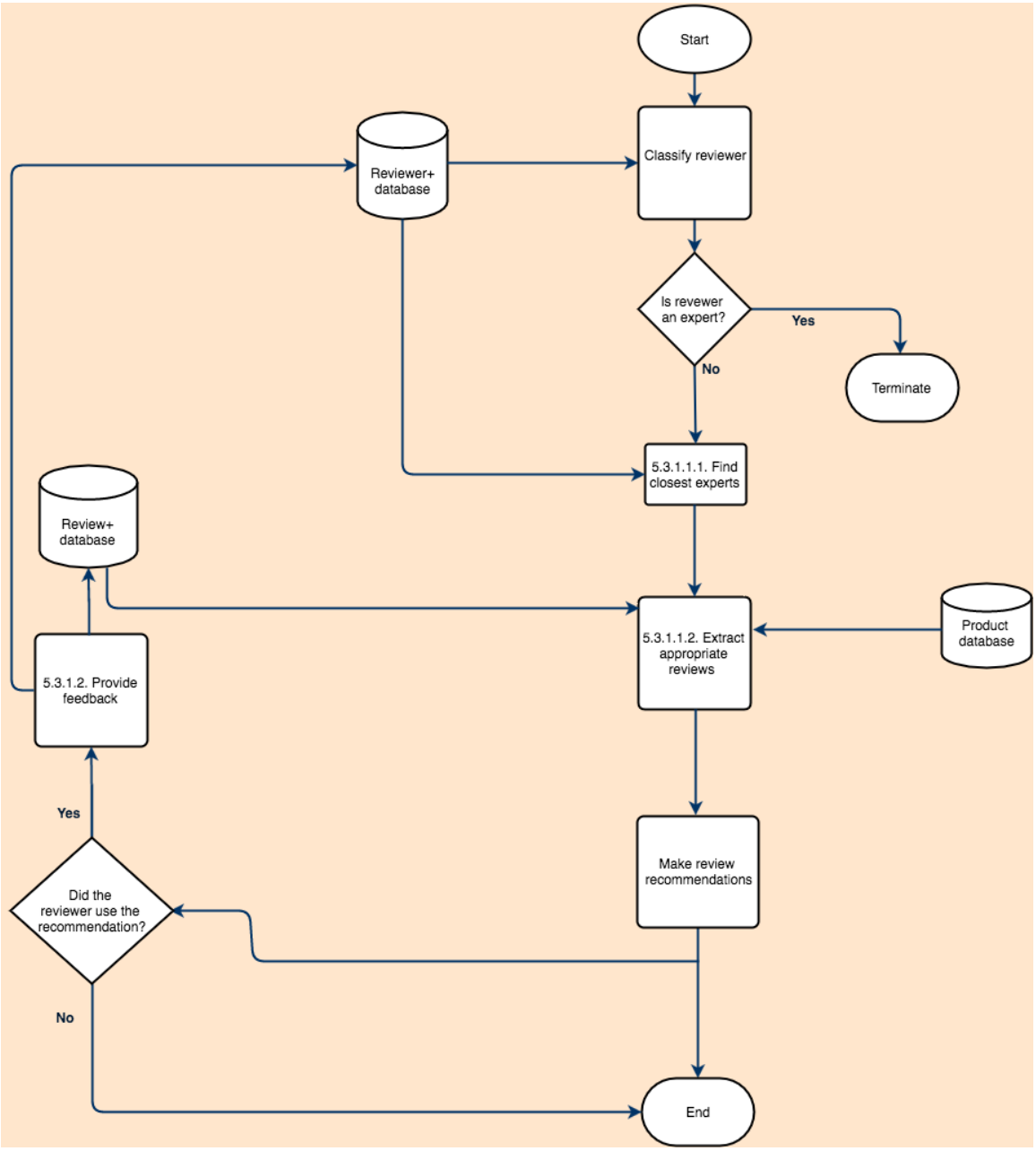

Figure 5.4: Flowchart of recommendation engine and feedback process

In Figure 5.4, we can see that there are multiple processes and decisions that are made in order to generate recommendations (covered in Section 5.3.1.1) followed by the feedback process (covered in Section 5.3.1.2).

5.3.1.1 Recommendation generation

Recommendation generation is the process of generating suitable recommendations for a given reviewer from the list of past reviews posted by experts. Based on the flowchart presented in Figure 5.4, below is the algorithm to generate recommendations for a reviewer *r* to review product *p*:

Algorithm RecommendationsGeneration (*r*, *p*)
Inputs: Reviewer *r*; Product *p*
Database used: Reviewer+ database
Review+ database
Product database
Returns: List of reviews to be recommended to *r*

1. *appropriateReviewList* ← []
2. **If Classify**(*r*) does *not* **return** "expert" **then** // either conscientious or novice
3. *closestExpertList* ← **FindClosestExpert** (*r*, *k*) // find *k* closest experts
4. *appropriateReviewList* ← **ExtractAppropriateReviews** (*r*, *p*, *closestExpertList*))
// appropriate reviews based on *p*
// and past reviews of *closestExpertList*
5. **Endif**
6. **Return** *appropriateReviewList*

End Algorithm

The first step is to classify reviewer *r* into one of the classes—expert, conscientious or novice. The process of classification is covered in detail in Section 4.4 of Chapter 4. The reviewer+ database contains the class label of each reviewer along with other information about the reviewer. **Classify**(*r*) returns the class label from the reviewer+ database. The process of recommendation generation is carried out *only* if the reviewer *r* is not an expert. There are two main function calls **FindClosestExperts()** and **ExtractAppropriateReviews()** that correspond to two modules, respectively: ((1) Finding closest experts for a reviewer, and (2) Extracting appropriate reviews.

Finding closest experts for a reviewer is the first step, which mainly focuses on searching the experts closest to the reviewer for who recommendations are being

generated. The closeness of a reviewer with experts can be measured using a distance metric. We will cover the details of finding a reviewer's closest experts in Section 5.3.1.1.1. The Second step is to extract appropriate reviews from the list of past reviews posted by the group of closest experts based on product type. This step mainly involves utilizing the findings we made about the relation of helpfulness with review features such as review length or review overall (ratings) based on product type. We will discuss extraction of appropriate reviews in detail in Section 5.3.1.1.2.

5.3.1.1.1 Find closest experts

Traditionally, in recommendation systems, memory-based methods have been used to compute similarities between users, referred to as ***user-based collaborative filtering*** (Sarwar *et al*., 2001). Basically, memory-based algorithms are heuristics that make rating predictions based on the entire collection of previously rated items by the users to recommend the most suitable items (Adomavicius and Tuzhilin, 2005). The similarity between two users is computed based on their ratings of common items. User-user similarity ranges from 1 if they are totally similar, and -1 if they are completely dissimilar. Based on the rating of the most similar users, it predicts the rating the current user would give to every item he or she has not rated yet. Then the recommendation system suggests the item with the highest rating to the user (Massa *et al*., 2004). Since our framework aims at improving the reviewing skill of reviewers, we focus on finding reviewer-reviewer similarity, instead, based on the reviewers' evolution-trend.

Recall that, from Section 5.1, we have established that reviewers evolve over time by improving their reviewing skill. However, the rate of improvement may differ from

one reviewer to another. Reviewer similarity is calculated based on the similarity between their evolution rates.

Also recall that, from Chapter 4, we know that, in each product categories, there are reviewers who have expertise in writing a good quality review, identified as *expert* reviewers. ***The recommendation framework helps a given reviewer r to improve their reviewing skill by recommending reviews from k closest experts***. To improve the reviewing skill of reviewers with lesser reviewing expertise, we calculate conscientious-expert similarity or novice- expert similarity to compute their closeness. The closeness of a reviewer with experts is measured in terms of the closeness of their review quality over time with one another. We use the *Euclidean distance* between the slopes of average helpfulness of the given reviewer *r* and experts on a yearly basis. The *k* experts with the least distances with the reviewer *r* are labeled as *k* closest expert of, where *k* is a user-defined value, which denotes the number of closest experts. For example, Table 5.10 contains the average helpfulness of a random conscientious reviewer and its 3 closest experts (i.e., $k = 3$), along with their respective *Euclidean distance* denoted by *d* with the conscientious reviewer in Books category for 5 active years.

	Average Helpfulness			
Year count	**Conscientious Reviewer** $d = 0.000$	**Expert 1** $d = 0.013$	**Expert 2** $d = 0.016$	**Expert 3** $d = 0.017$
1	0.833	0.857	1	0
2	0	0	0	0.667
3	0.667	0.667	0.667	0.921
4	0.916	0.95	0.875	0.925
5	1	1	1	0.928

Table 5.10: Average helpfulness of a random conscientious reviewer and three closest experts in "Books" category.

In Table 5.10, the average helpfulness of the conscientious reviewer decreases from Year 1 to Year 2 and then increases gradually from then to Year 5. The three closest

experts to this reviewer also have the similar change in their average helpfulness over a period of five years. Since the evolutionary trend of the closest experts is similar to that of the reviewer, our recommendation system suggests the reviews posted by the closest experts. In other words, if a reviewer evolves her review quality in a similar trend as her closest experts for active year *t* then the reviews posted by the closest experts in the active year *t*+1 can be recommended to the reviewer in the same year *t*+1. For a review to become *recommendable*, its helpfulness must be greater than the average *helpfulness* at time *t* of the reviewer. If the review's helpfulness is less than the average *helpfulness* of the reviewer, then the review is *non-recommendable*.

In Table 5.11 there are 7 random conscientious reviewers, their average helpfulness and the helpfulness of three different reviews found for each reviewer. These reviews were posted by three different closest experts in the order of increasing distance denoted by *d* (Review 1 is posted by the expert with the shortest distance and Review 3 is posted by the expert with the longest distance).

	Average Helpfulness			
Conscientious Reviewer id	**Conscientious Reviewer**	**Review 1**	**Review 2**	**Review 3**
Reviewer 1	0	0.75 (d = 0.001)	0.571 (d = 0.004)	0.6 (d = 0.005)
Reviewer 2	0.861	0.875 (d = 0.005)	0.875 (d = 0.007)	**0.8** (d = 0.010)
Reviewer 3	1	1 (d = 0.006)	1 (d = 0.006)	1 (d = 0.007)
Reviewer 4	0.682	1 (d = 0.006)	0.768 (d = 0.011)	0.889 (d = 0.017)
Reviewer 5	0.716	0.871 (d = 0.046)	0.889 (d = 0.0053)	**0.633** (d = 0.068)
Reviewer 6	0.25	0.333 (d = 0.006)	1 (d = 0.007)	0.75 (d = 0.007)
Reviewer 7	1	1 (d = 0.005)	1 (d = 0.007)	**0.928** (d = 0.010)

Table 5.11: Average helpfulness of 7 random conscientious reviewers and helpfulness of three reviews posted by three closest experts in "Books" category. Bolded values indicate non-recommendable reviews.

In Figure 5.11, we see that it is possible for the helpfulness of some reviews (bolded) posted by closest experts to be less than the average helpfulness of the reviewer as the distance *d* increases. However, most of the reviews (non-bolded) posted by closest experts have helpfulness greater than the average helpfulness of the reviewer. As observed in Table 5.11, the generated reviews may be *non-recommendable* as distance increases. Therefore, we check the sensitivity of *k* for each product category to determine the appropriate values of *k* such that the reviews generated by the closest experts are likely to be *recommendable*. The detail on setting the value of *k* is covered in Section 5.3.1.1.1.1.

The algorithm to find *k* closest experts for a reviewer *r* is below. The algorithm uses the data structure of reviewer whose definition is covered in Appendix, Section G.

Algorithm FindClosestExperts(*r*, *k*)
Inputs: Reviewer *r*; Integer *k*
Database used: Reviewer+ database
Review database
Returns: List of *k* closest expert reviewers to *r*

1. *closestExpertList* ← []
2. *r.avgHelpfulnessList* ← list of average helpfulness of *r* in each active year
3. **For** each expert *e*
4. *e.avgHelpfulnessList* ← list of average helpfulness of *e* in each active year
5. *d* ← **CalculateEuclideanDist** (*r*, *e*) //Euclidean distance
5. add *d* to *expertDistanceList*
6. **Endfor**
7. **sort** *expertDistanceList* in ascending order
8. *closestExpertList* ← experts corresponding to top *k* distances in *expertDistanceList* // closest experts have shortest distance
9. *r. prevAvgHelpfulness* ← *r. avgHelpfulness* // capture the avg helpfulness of *r* before // recommending
10. **Return** *closestExpertList*

End Algorithm

The algorithm to find *k* closest experts calls **CalculateEuclideanDist**(*r*, *e*) function that calculates Euclidean distance between reviewer *r* and an expert *e*. The Euclidean distance

is one of the most common distances that have been used for numerical data (Gan *et al*., 2007). We used this distance to measure the distance between the average helpfulness between any two reviewers r_i and r_j. For a reviewer r_i, $r_i.avgHelpfulnessList = (r_{i,1}, r_{i,2}, ..., r_{i,n})$, and for another reviewer r_j, $r_j.avgHelpfulnessList = (r_{j,1}, r_{j,2}, ..., r_{j,n})$, *Euclidean distance* between r_i and r_j is defined as,

$$d(r_i, r_j) = \sqrt[2]{\sum_{k=1}^{n} |r_{i,k} - r_{j,k}|^2}$$

where n is the number of active year of the reviewer r_i and $r_{i,k}$, $r_{j,k}$ are the values of average helpfulness for the kth active year of the reviewer r_i and r_j respectively.

5.3.1.1.1.1 Value of *k*

Our aim is to generate *recommendable* reviews i.e., reviews whose helpfulness is greater than the average helpfulness of the reviewer for whom recommendations are generated. As seen in Table 5.11 reviews tend to be *non-recommendable* as the distance between their author and the reviewer increases. We calculate the percentage of *non-recommendable* reviews with the increasing value of k in order determine the value of k where the *non-recommendable* reviews saturate. Figure 5.5 plots the percentage of non-recommendable reviews for different value of k ranging from 1 to 30 for products related with prevention consumption goal.

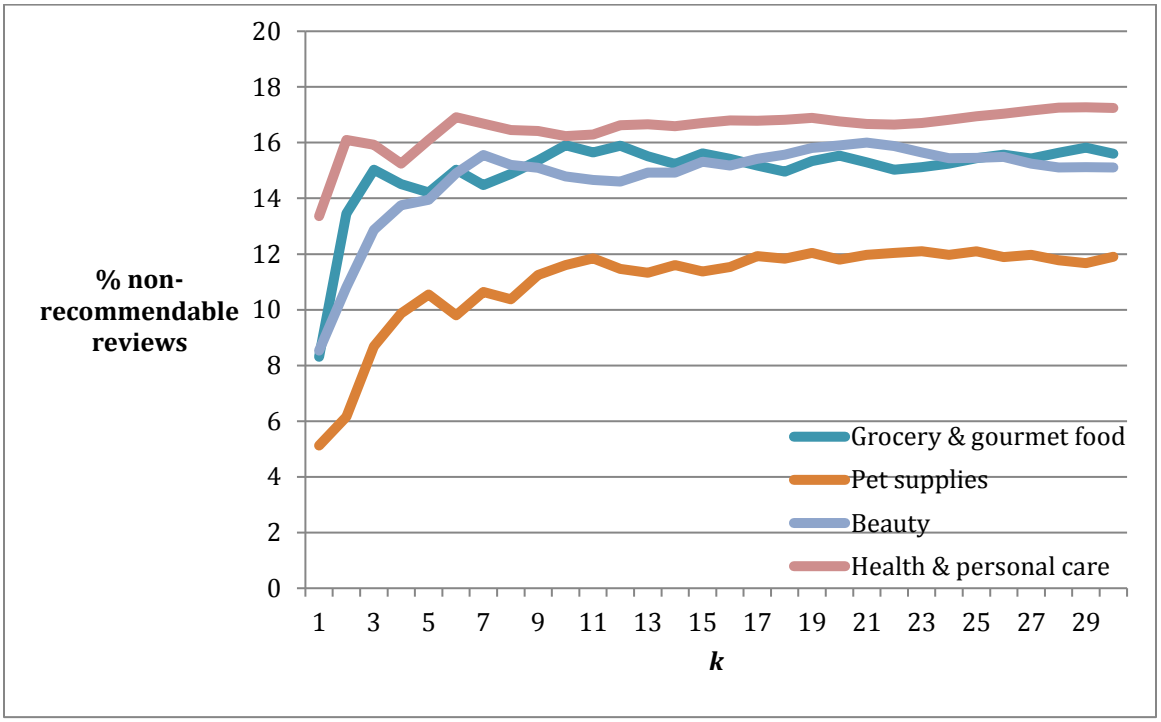

Figure 5.5: Percentage of non-recommendable reviews with respect to *k* for products related with prevention consumption goal.

In Figure 5.5, we see that the percentage of non-recommendable reviews among all the reviews found from the closest experts increases until a certain value of *k* after which it converges. For example, in Pet supplies the percentage of non-recommendable reviews increases until it converges at around 12% when the value of *k* reaches around 10. Figure 5.6 plots the percentage of non-recommendable reviews for different value of *k* ranging from 1 to 30 for products related with promotion consumption goal.

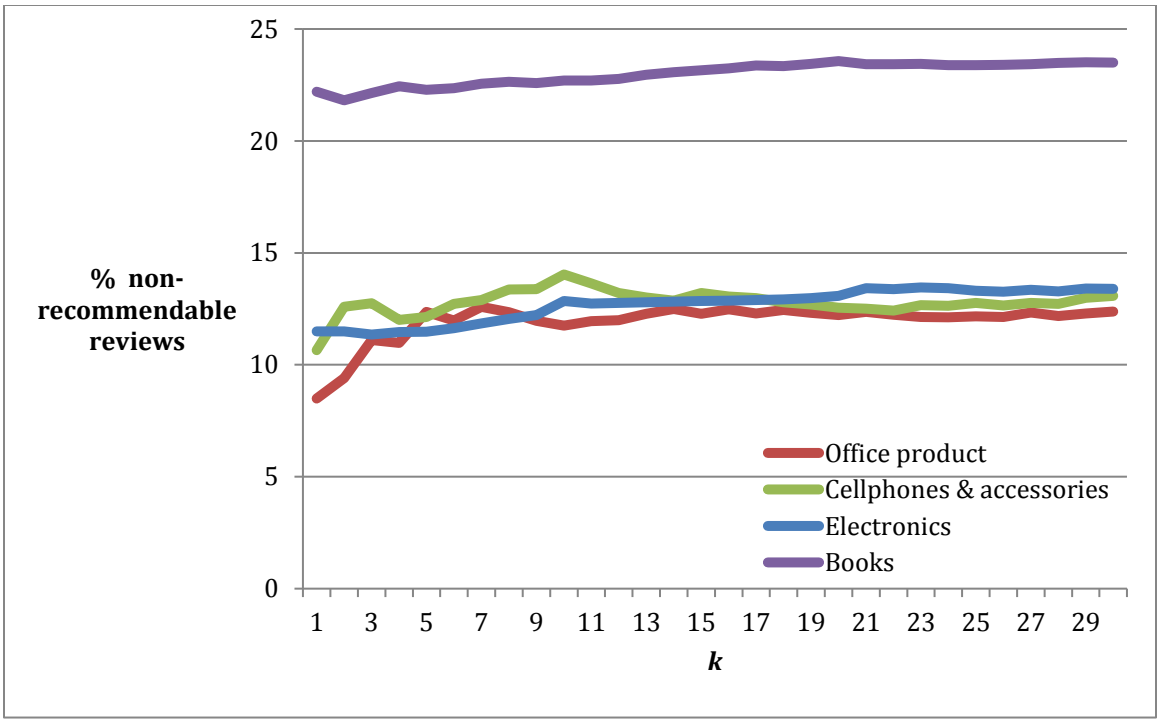

Figure 5.6: Percentage of non-recommendable reviews with respect to *k* for products related with promotion consumption goal.

In Figure 5.6, similar to Figure 5.5, the percentage of non-recommendable reviews increases until a certain value of *k* after which it converges. For example, in Books and Electronics, the percentages slowly increase until they converge at 23% and 13%, respectively when *k* reaches around 20. To summarize Figures 5.5 and 5.6, the percentages of non-recommendable reviews and range of *k* at convergence point are listed in Table 5.12 for all product categories.

Category	**Range of k at convergence point**	**% Non-recommendable reviews at convergence**
Books	19-21	23
Electronics	19-21	13
Cell Phones and accessories	9-11	13
Office product	7-9	12
Grocery and gourmet food	9-11	16
Health and personal care	7-9	16
Beauty	7-9	16
Pet supplies	9-11	12

Table 5.12: Range of *k* where percentage of non-recommendable reviews converges

In Table 5.12, for all product categories except Books and Electronics, the convergence occurs when *k* ranges from 7 to 11. So for our framework, an appropriate value of *k* would range from 7 to 11. However, for Books and Electronics, the convergence occurs late when *k* has reached around 19- 21. Note from Table 4.1 in Chapter 4, the product count for Books (~1 million) and Electronics (~300k) is the highest and second highest compared to the other categories. These two product categories are very diverse and thus have more diverse reviews in each category. Subcategorization of products (and their reviews) may decrease the convergence point in these cases. For example, Books may be subcategorized into drama, science fiction, horror, mystery, romance, action, etc. and Electronics may be subcategorized as audio & video, camera & photo, car electronics, computer, etc. This needs further investigation and could be one of the interesting future works.

5.3.1.1.2 Extract appropriate reviews

We matched the evolution trend of a reviewer with experts for time *t* to find a group of closest experts. Our assumption is that a reviewer will reach the highest level of expertise by learning from the experiences of his or her closest experts. In order to provide assistance with intermediate steps for the reviewer to grow from his or her current state to the highest level of expertise, we want to recommend the positive actions—that are not too far out of reach of the reviewer—and discourage the negative actions—that are within reach of the reviewer—of the reviewer's closest experts. The list of reviews posted by the closest experts at time *t* +1 can either increase or decrease their average helpfulness. *Our recommendation system framework recommends the reviews in 'recommendable list' that will likely improve the average helpfulness along with the reviews in 'non-*

recommendable list' that will likely decrease the average helpfulness of the reviewer for who the recommendations are generated. Therefore, we segregate the list of reviews posted by closest experts into *recommendable* and *non-recommendable* reviews. *Recommendable* reviews are the reviews whose helpfulness is greater than the average helpfulness of the reviewer for whom the recommendations are generated. *Non-recommendable* reviews are the reviews whose helpfulness is less than the average helpfulness of the reviewer for whom the recommendations are generated. We know from Figures 9 and 10 that percentage of *non-recommendable* reviews is very less (usually <16%) compared to *recommendable* reviews for most product categories. If the reviewer uses any of the *recommendable* reviews to write her new review, our premise is that we expect, as a result, that *the helpfulness of the new review will be proportional to the recommended review*. Hence the newly posted review is likely to increase the average helpfulness of the reviewer. The *non-recommendable* reviews, also posted by the closest experts at time t+1, can be used to warn the reviewer on how not to write reviews. This could be very useful to the reviewer, as it would help them to avoid the reviews they would have posted otherwise and not repeat the same mistakes their closest experts made.

Once the list of *recommendable* reviews is created, these reviews are prioritized based on the conclusions on product category we derived in Table 4.34 in Chapter 4. To restate what was mentioned in Chapter 4, how a review is helpful is based on the type of product reviewed and the associated review length and overall/ratings.

- Higher overall leads to helpful reviews for products associated with promotion consumption goal such as Books, Electronics, Cell Phones and accessories, and Office product.

- Longer review leads to helpful reviews for products associated with prevention consumption goals such as Grocery and gourmet food, Health and personal care, Baby, Beauty, and Pet supplies.

Based on the above findings, the past reviews of experts are prioritized before recommending to the reviewer. For reviews related to promotion consumption goal products, reviews with high overall/rating are prioritized over lower ones. For reviews related to prevention consumption goal products, longer reviews are prioritized over shorter reviews. Figure 5.7 shows the flowchart of the extraction process.

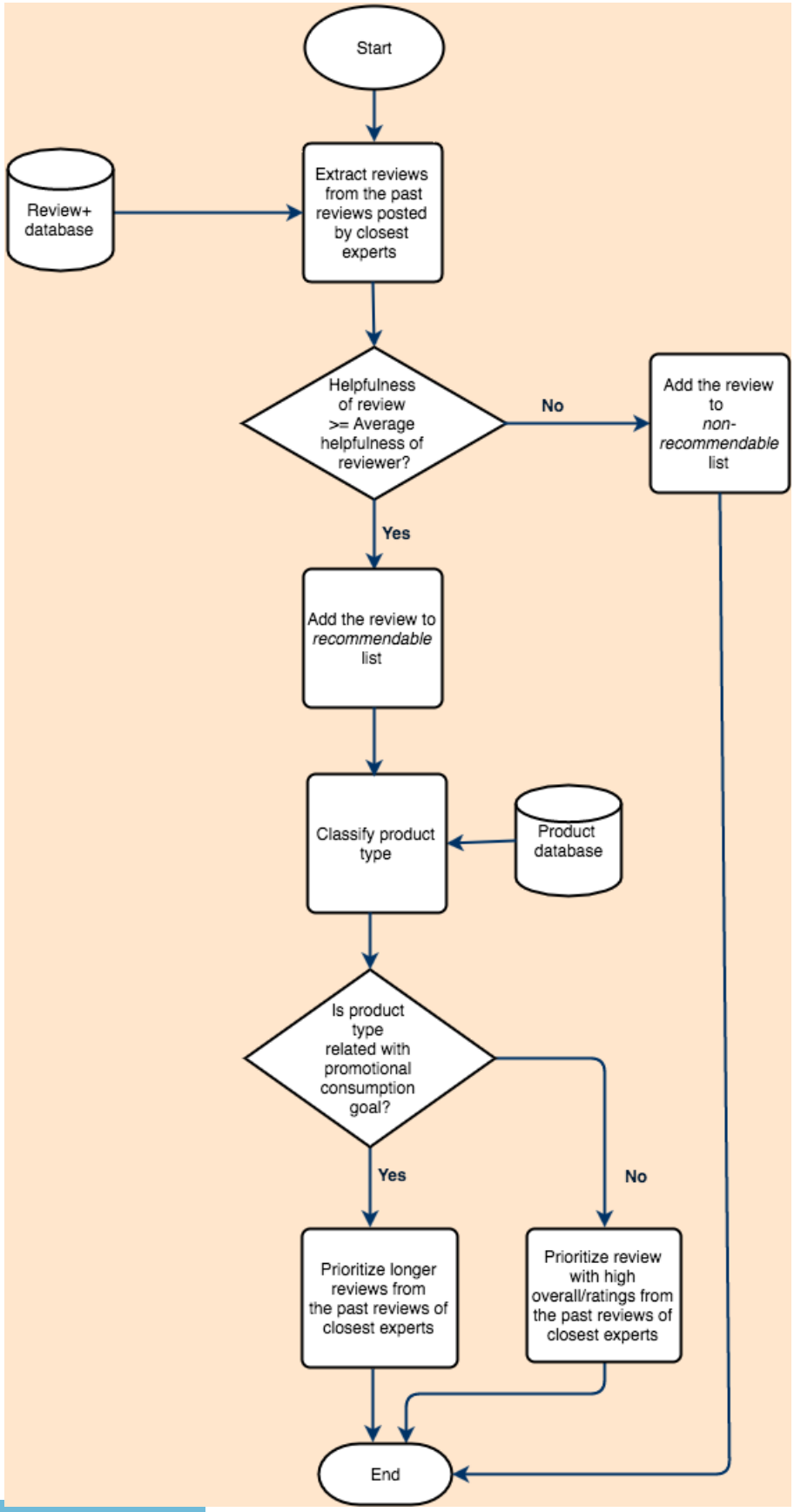

Figure 5.7: Flow chart to extract appropriate reviews

The algorithm to extract appropriate reviews is below. The algorithm uses the data structure of review and reviewer whose definitions are covered in Appendix, Section G.

Algorithm ExtractAppropriateReviews (*r*, *p*, *closestExpertList*)
Inputs: Reviewer *r*; Product *p*; List of closest experts *closestExpertList*
Database used: Product database
Review+ database
Returns: List of reviews to be recommended and avoided

1. *recommendableList* ← []
2. *nonRecommendableList* ← []
3. **For** each expert *e* in *closestExpertList*
4. add reviews posted by *e* to *recommendableList*
5. **Endfor**
6. **For** each review *rev* in *recommendableList*
8. **If** *rev.helpfulness* < *r. prevAvgHelpfulness* **then** // non-recommendable reviews
9. move *rev* from *recommendableList* to *nonRecommendableList*
10. **Endif**
11. **Endfor**
12. **If** *p.type* is "prevention consumption goal" **then**
13. **sort** *recommendableList* in descending order of their review length
14. **Else** // *p.type* is "promotion consumption goal"
15. **sort** *recommendableList* in descending order of their overall/ rating
16. **Endif**
17. **Return** *recommendableList*, *nonRecommendableList*

End Algorithm

In the above algorithm product category is used to prioritize the *recommendable* reviews of closest experts. In the end of this phase, lists of *recommendable* and *non-recommendable* reviews are generated. Additionally, there are multiple other ways to prioritize *recommendable* reviews for recommendation, which can be integrated in the above algorithm. Below are some of the methods:

1. Use ***history of the reviews*** to prioritize *recommendable* reviews. The history of a review can provide information on how it was received, if it was recommended in the past. If the review was recommended in the past and was successful then it is likely to be successful again. Hence the review will have higher priority. There

could be multiple scenarios such as— reviews with no history of past recommendation, reviews with the history of *successful* recommendation, and reviews with the history of *unsuccessful* recommendation. To restate what was mentioned in Section 5.3, the recommended reviews that increased the average helpfulness of reviewer (for whom recommendations were generated) are referred as *successful recommendations* whereas those recommendations that did not increase the average helpfulness of reviewer (for whom recommendations were generated) are referred as *unsuccessful recommendations*. If a review has the history of past recommendation then its usage history such as *score* and *recommendation count* is used to calculate its *rank* that denotes its success rate till date (covered in detail in Section 5.3.1.2). The *score* of a review denotes the number of times the review has been *successful* and *unsuccessful*. The *score* is computed by adding 1 for each *successful recommendation* and subtracting 1 for each *unsuccessful recommendation*. Additionally, the number of times the review has been recommended in the past is stored as *recommendation count*. *Rank* denotes the success rate of a review and is calculated by dividing the *score* of the review with the *recommendation count* of the review. A review with high *rank* denotes high success rate. The code snippet to compute *rank* of a review is below:

```
1. For each review rev in recommendationList
2.       rev.rank ← rev.score / rev.recommendationCount
                                        // successful reviews have higher rank
3. Endfor
```

Once the *rank* of all reviews is calculated, the reviews are prioritized based on the distance between reviewer and expert (reviews' author), and their *rank*. The equation to calculate the *priority* of each review is below:

$$priority_{rev} = \frac{1 + rank_{rev}}{d(r, rev.author)}$$

where $d > 0$ and $-1 \geq rank \geq 1$. d is the distance between reviewer r (for whom recommendations are generated) and author of the review; and *rank* is the success rate of the review *rev*. The *priority* of a review is inversely proportional to the distance between the review's author and the reviewer (for whom recommendations are generated) because from Section 5.3.1.1.1.1 we know that the percentage of *non-recommendable* reviews increases as distance increases. In other words, the reviews posted by the expert with shorter distance will have higher *priority* than the reviews posted by the expert with longer distance. Also, the *priority* of a review is directly proportional to its success rate or *rank*.

2. Use ***history of expert reviewers*** to prioritize *recommendable* reviews. The history of an expert reviewer can provide information on how his or her reviews were received, if they were recommended in the past. If the reviews written by the expert were recommended in the past and were successful then it is likely that the reviews posted by the expert will be successful. Hence reviews posted by the expert will have higher priority. Similar to reviews, there could be multiple scenarios with reviewers such as— reviewers with no history of past recommendation, reviewers with the history of *successful* recommendation, and reviewers with the history of *unsuccessful* recommendation. The *rank* of the reviewers is calculated in a similar manner as the reviews. *Rank* of a reviewer denotes the success rate of the reviewer (covered in Section 5.3.1.2). In short, a reviewer whose reviews have a high success rate of recommendation has a high *rank*. The code snippet to calculate the *rank* of all reviewers is below:

```
1. For each review rev in recommendationList
2.     e ← rev.author                    // expert reviewer e who posted review rev
3.     e.rank ← e.score / e. recommendationCount
                                         // successful reviewers have higher rank
4. Endfor
```

Once the *rank* of all reviewers is calculated, the reviews can be prioritized based on the distance, and *rank* of the reviewer. The equation to calculate the *priority* of each review is below:

$$priority_{rev} = \frac{rank_{rev.author}}{d(r, rev.author)}$$

where $d > 0$ and $-1 \geq rank \geq 1$. *d* is the distance; and *rank* is the success rate of the author. The reviews in recommendation list are sorted based on their *priority* value before being displayed to the reviewer. The *priority* of a review is directly proportional to its success rate or *rank* of the review author. In other words, if an author has a history of successful recommendations in the past then a review posted by the author is likely to be *successful* hence the review has a higher *priority*.

3. Use ***history of reviews and reviewers*** to prioritize the *recommendable* reviews. This approach is a weighted combination of the above two approaches. We use the *rank* of review and its author to calculate the *priority* of each review as written below:

$$priority_{rev} = \frac{w_1 * rank_{rev} + w_2 * rank_{rev.author}}{d(r, rev.author)}$$

$$w_1 + w_2 = 1$$

where *d* ($d > 0$) is the distance. The reviews in recommendation list are sorted based on their *priority* value before being displayed to the reviewer.

5.3.1.2 Feedback process

The final component in recommendation system framework is feedback process. This provides feedback on both review and reviewer (author of the review) based on the usage of recommended reviews. The feedback process on any recommended review starts after a certain duration of time (say 10 weeks). This duration is the time when other customers read the posted review (recommended by our framework) and post their votes, which determines the helpfulness of the review. The feedback is used to record the usages of recommended reviews for future use. If a recommended review was used by a reviewer and successfully increased the average helpfulness of the reviewer, it is referred as a *successful recommendation*. If the recommended review fails to increase the average helpfulness of the reviewer then it is referred as an *unsuccessful recommendation*. At the end of this phase, the review+ database is updated to track review usage data such as *score* and *recommendation count.* The *score* denotes the number of times the review has been *successful* or *unsuccessful.* For each successful recommendation 1 is added whereas for each unsuccessful recommendation 1 is subtracted to compute the *score* of the recommended review. The number of times a review has been recommended in the past is stored in *recommendation count*.

Additionally, feedback process also provides feedback on the expert reviewer who wrote the recommended review (author of the review). It updates reviewer+ database to store the *score* and *recommendation count* of the expert. Similar to a review, the *score* of an expert reviewer is calculated by adding 1 for his or her each *successful recommendation* and subtracting 1 for each *unsuccessful recommendation.*

Recommendation count of an expert reviewer denotes the number of times his or her review has been recommended. The algorithm of feedback process is below:

Algorithm FeedbackProcess (*rev*, *r*)
Inputs: Recommended review *rev*; Reviewer *r*
Database used: Review+ database
Reviewer+ database
Returns: status of *rev*

1. *e* ← *rev.author*
2. **If** *rev.helpfulness* <= *r. previousAvgHelpfulness* **then** // unsuccessful review
3. *status* ← unsuccessful
4. *rev.score* ← *rev.score* -1
5. *e.score* ← *e.score* -1
6. **Else** // successful review
7. *status* ← successful
8. *rev.score* ← *rev.score* +1
9. *e.score* ← *e.score* +1
10. **Endif**
11. *rev.recommendationCount* ← *rev.recommendationCount* +1
12. *e.recommendationCount* ← *e.recommendationCount* +1
13. **Return** *status*

End Algorithm

As stated earlier in Section 5.3.1.1.2, *score* and *recommendation count* measure the success rate of a review, denoted as *rank*. Reviews with higher *rank* get prioritized as templates for future use and reviews with lower *rank* get discouraged from future use. However, lower *rank* reviews may also be used to warn reviewers on how **not** to review a product. Similarly, for a reviewer *rank* denotes the success rate of the reviewer. A reviewer with a higher *rank* is more reliable as his or her reviews have the higher success rate. Similarly, the reviews posted by the expert reviewer with higher *rank* are encouraged over the reviews of the reviewer with lower *rank*.

5.3.2 Mockup diagram

In this section, we present mockup diagrams of the proposed recommendation software. This prototype implements review recommendation system framework focusing mainly

on recommendation presentations and user interactions with the recommended reviews. In this section, we use the mockup diagrams to showcase how we can present review recommendations and how reviewers may use *recommendable* reviews as well as *non-recommendable* reviews. Figure 5.8 shows how a reviewer interacts with the software by reading through *recommendable* reviews (Figures 5.10 and 5.11) and *non-recommendable* reviews (Figure 5.12).

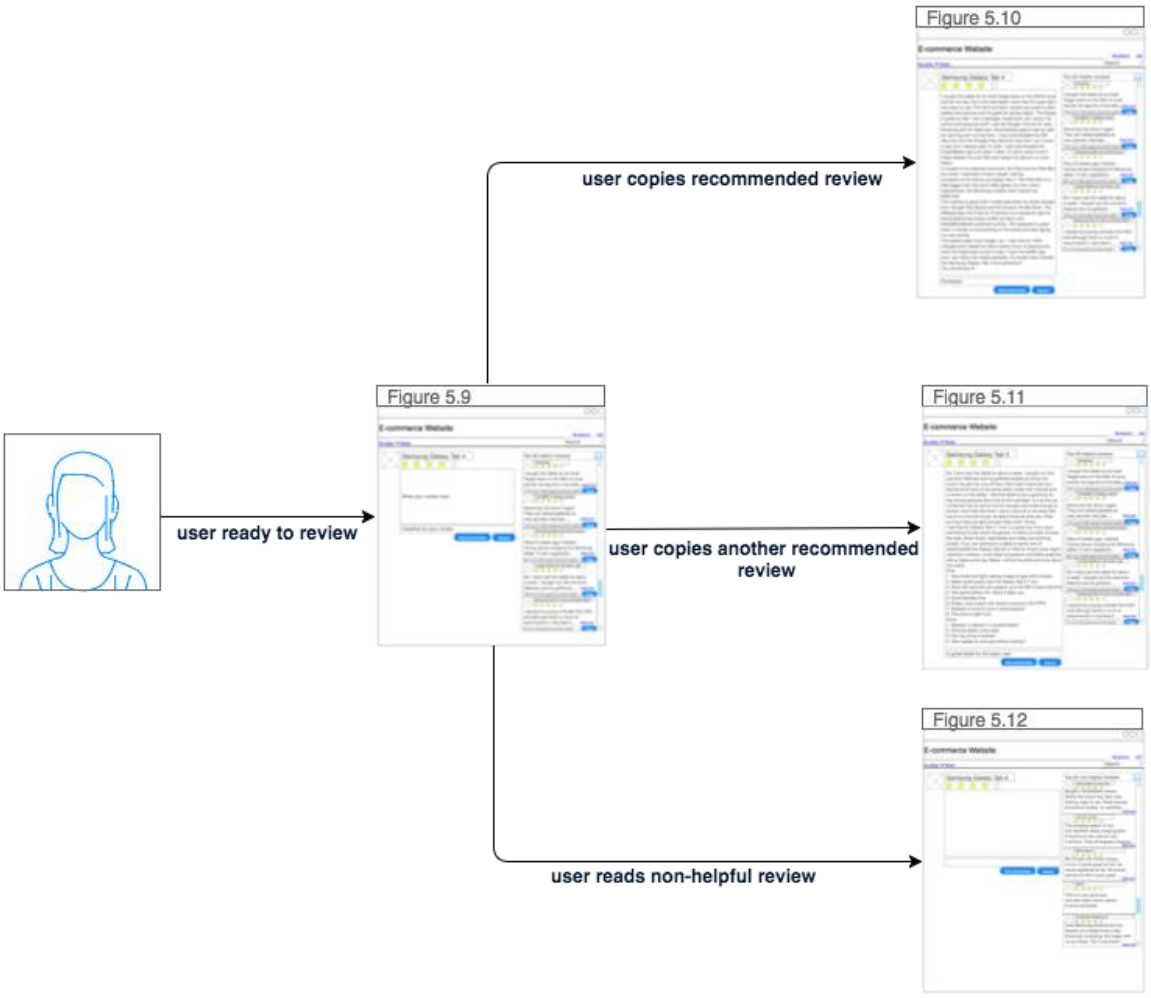

Figure 5.8: User interaction with the system making use of *recommendable* and *non-recommendable* reviews

We will now discuss how Figures 5.9, 5.10, 5.11 and 5.12 look and how a user can make use of them. Figure 5.9 shows a user interface that recommends a list of reviews to a reviewer who is ready to write a review for an electronics product Samsung Galaxy Tab

4. There are five reviews in the right panel, which is scrollable and contains reviews from *k* closest experts. The reviewer may use any one of the recommended reviews to write his or her own review.

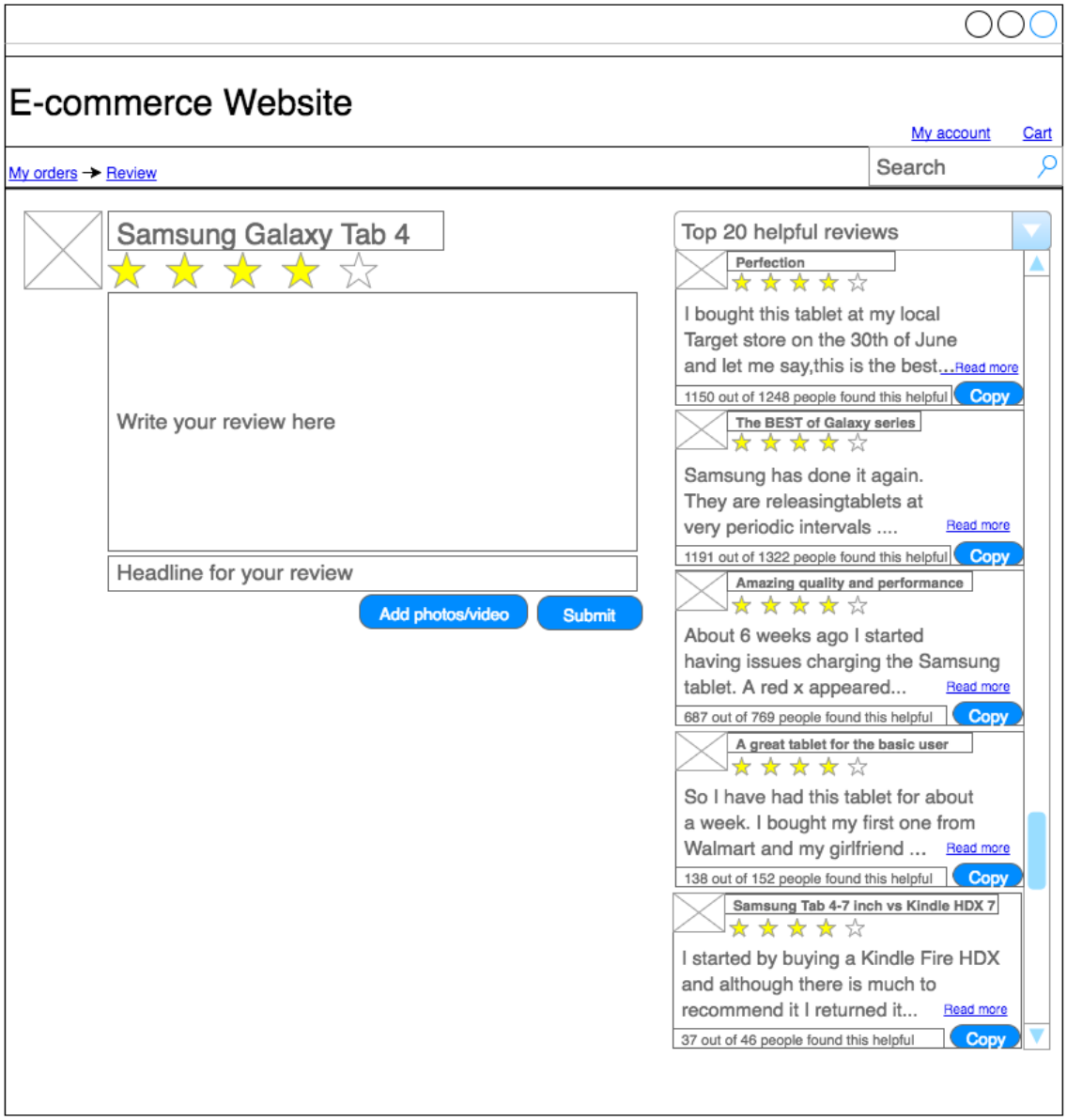

Figure 5.9: : Interface displaying top 20 helpful reviews in the *recommendable list* in right panel

In Figure 5.9, the right panel contains "Copy" button for each recommended review. The "Copy" button is clicked to copy the respective review into the review description box. Figure 5.10 shows the interface after the reviewer has clicked the first (topmost) "Copy" button.

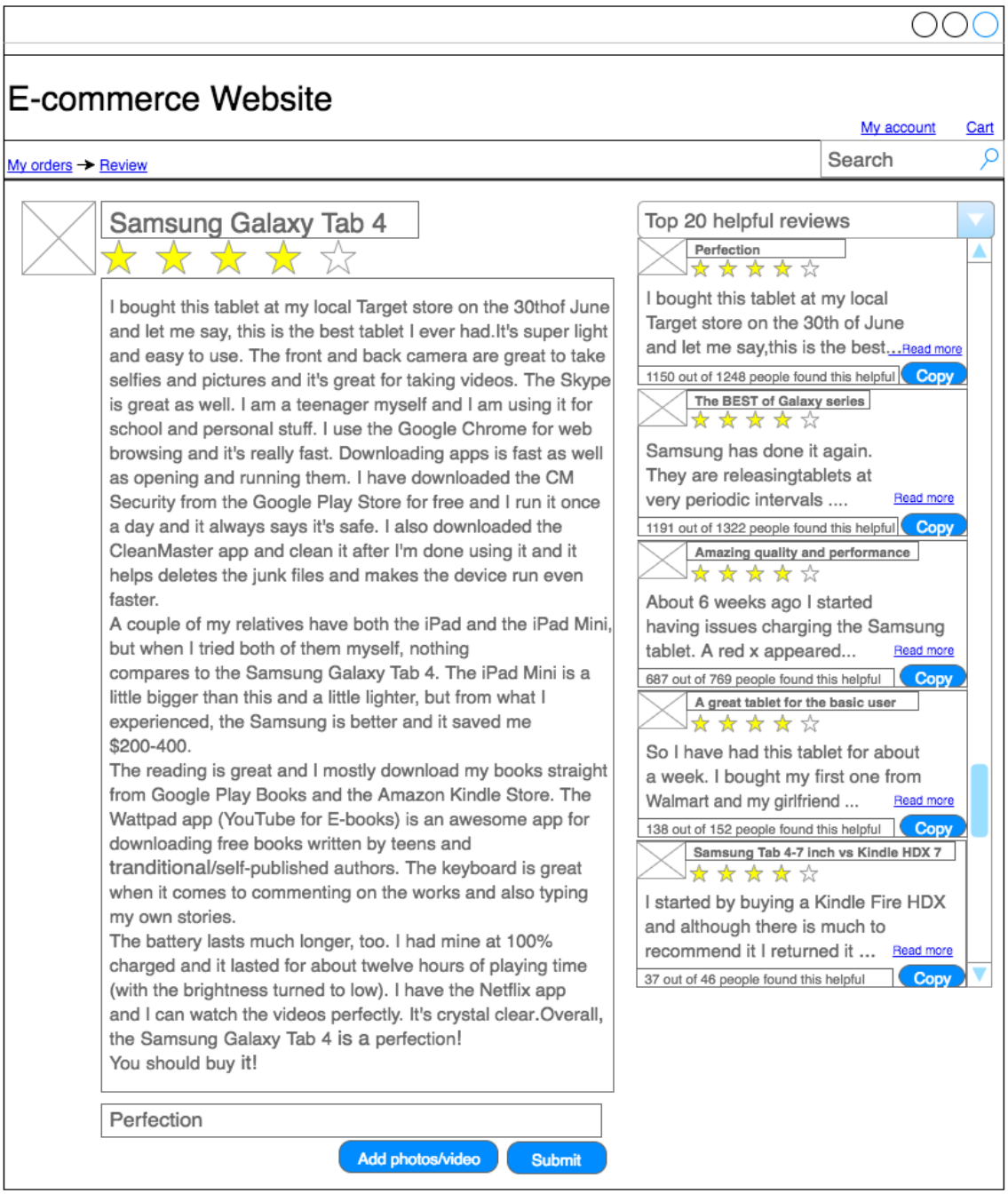

Figure 5.10: User interface after the first copy button is clicked i.e., the first review in the *recommendable list* is copied to the review description textbox.

Copy button copies the recommended review headline and description into the description textbox and headline textbox respectively. Similarly, Figure 5.11 below is another example of the interface after the reviewer has clicked the fourth "Copy" button.

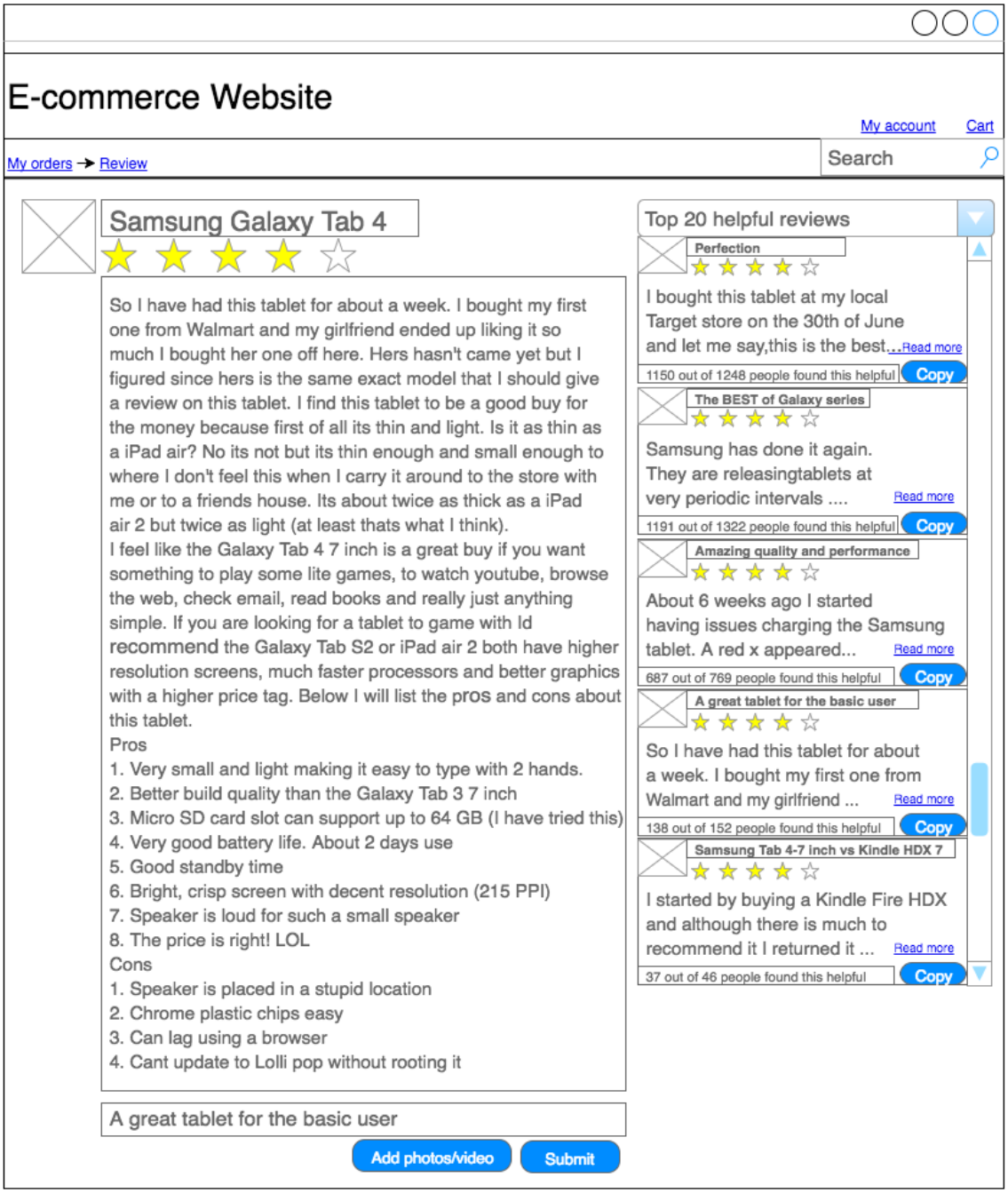

Figure 5.11: User interface after the fourth copy button is clicked i.e., the fourth review in the *recommendable list* is copied to the review description textbox.

Once the recommended review has been copied, the reviewer may choose to edit the review description and headline before posting it. Similar to reviews in the *recommendable list*, the reviews in the *non-recommendable list* may be displayed to warn reviewers on how *not* to write a review. Below Figure 5.12 shows the interface of top 20 non-helpful reviews.

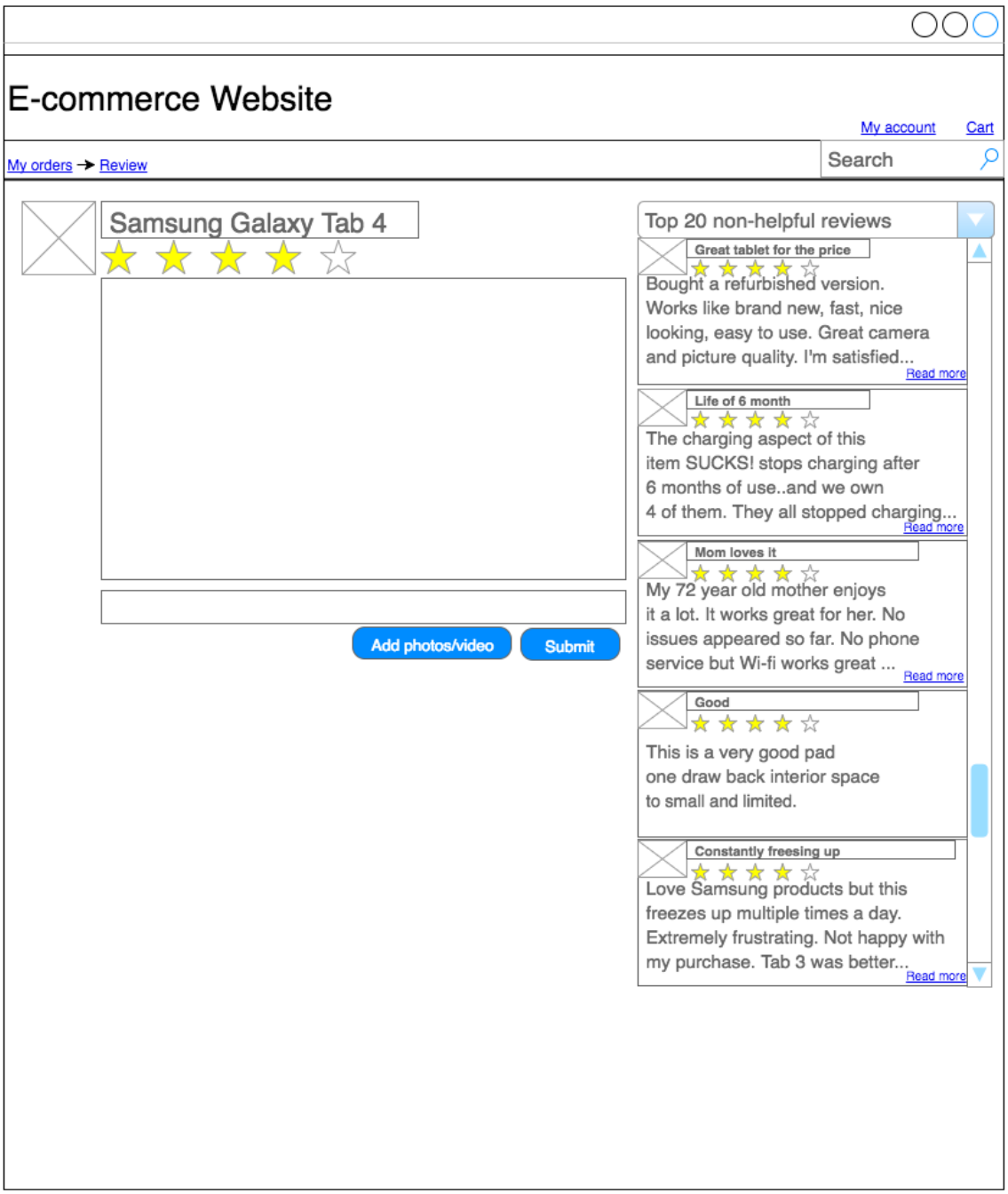

Figure 5.12: Interface displaying top 20 non-helpful reviews in the *non-recommendable list* in right panel

Figure 5.12 does not have "Copy" button i.e., reviewers cannot copy non-helpful reviews but they can read through all of them.

The review recommendation software should be able to track how reviewers used the *recommendable* and *non-recommendable* reviews in order to better understand the user

preferences. For example, if reviewers edited the recommended review before posting or not. This could be an interesting future work and we will cover it in detail in Chapter 6.

Chapter 6
Conclusions & Future Work

In this chapter, we summarize the findings of this thesis and then overview ideas for future work. Section 6.1 will review the conclusions and Section 6.2 will present ideas for future work.

6.1 Conclusions

In this Thesis, we investigated an Amazon.com database of 2.3 million reviewers to understand their reviewing skills and how those skills changed over time. Their reviewing skill was observed based on their reviews in nine different product categories such as Books, Electronics, Cellphones and accessories, Grocery and gourmet food, Office product, Health and personal care, Baby, Beauty, and Pet supplies. We then proposed a review recommendation framework to train reviewers to better write about their experiences with the product by leveraging the behaviors of expert reviewers who are good at writing helpful reviews. Specifically, in Chapter 4, we used *X*-means clustering technique to model reviewers into different classes based on their review quality. Also in Chapter 4, we used classification approaches to investigate how reviews are perceived differently across different product categories. We then analyzed how different classes of reviewer evolve over time in regard to their reviewing skill in Chapter 5. Also in Chapter 5, we proposed review recommendation system framework that is based on the reviewer evolution to generate review recommendations that help train

reviewer to write better quality review. We can make several conclusions based on the following key findings of this Thesis:

Finding 1. Reviewers have different skill levels for posting quality reviews. Reviewers may differ in their expertise level such as expert, novice, etc. based on the quality of their reviews. The expertise levels reflect the ability of a reviewer to write quality review, which is measured by the number of up-votes the review receives from customers. Amazon uses "helpfulness" as the primary way of measuring consumers' evaluation of a review. Therefore, expertise level of each reviewer class is determined by their reviews *quality* i.e. helpfulness. Reviewers at different expertise level need different kind of training to write quality reviews. Therefore, recommendations are personalized to fit the expertise level of reviewers.

Finding 2. Reviews are valued differently across different product categories. Through machine learning based classification techniques, we identified the salient review features for different product categories. Using decision tree for classification we found the features that differentiated reviewer classes from one another. Understanding which review feature played important role to perform classification helped us find the features that are more important than other. As stated in Chapter 4, for products associated with *prevention* consumption goal such as Health and personal care, Grocery and gourmet food, Baby, Beauty, and Pet supplies *longer reviews are perceived to be more helpful*; and for products associated with *promotion* consumption goal such as Books, Cellphones and accessories, Electronics, and Office products *positive reviews are more helpful than negative ones*. These findings showed that reviews are perceived differently across

different product categories. We use this finding to make effective recommendations by generating product specific review recommendations.

Finding 3. Reviewers evolve over time. For some reviewers, evolution indicated improvement in their reviewing skills whereas for others it indicated the opposite. The actions performed by an expert reviewer can be recommended to reviewers with lower reviewing skills. We used this finding to *recommend the actions of expert cluster, which improves quickly to novice or* conscientious *cluster, which grow slowly or remain constant*. Based on reviewer evolution trend we proposed a review recommendation framework that can help a novice or conscientious reviewer to become an expert reviewer.

Review recommendation framework aims at improving the reviewing skill of a reviewer and therefore computes reviewer-reviewer similarity, based on the similarity between their evolution rates. To improve the reviewing skill of reviewers with lesser reviewing expertise, we calculate conscientious- expert similarity or novice- expert similarity to compute their closeness. The reviews posted by the closest experts are recommended to the reviewer. We verified that for a random conscientious reviewer, at least 80% of the reviews posted by closest experts were of higher quality than that of the conscientious reviewer. Therefore, **our recommendation system framework recommends the reviews that are of better quality than that of reviewer's**. Additionally, it will also warn the reviewer on how *not* to review by displaying the list of reviews that will likely decrease reviewer's skill. Hence the framework not only trains a reviewer on how to improve their review quality but also warns the reviewer on how to avoid the mistakes that may decrease their review quality.

6.2 Future Work

Future work can center on multiple facets:

- Designing recommendation software and data recording such that explicit feedback can be used to create future models.
- Enhancing reviewer-expert similarity based on elaborate reviewer and product profiling into the future version.
- Maturing recommendation framework by diversifying recommendations.

6.2.1 Designing recommendation software

In future work, we can develop software to implement the framework that we have proposed. Unlike traditional recommendation systems that suggest products such as movies, books, news, videos, and so on, the review recommendation system is very unique. Review recommendation software needs to be equipped with displaying the recommendations in a user-friendly manner such that reviewer feel motivated to read through the list of recommendations before settling to use one of them. Section 5.3.2 in Chapter 5 presents some ideas for designing the software. There are multiple ways a reviewer could use the recommended reviews such as 1) post the recommended review as it is, 2) make minor changes to the recommended review before posting, 3) make use of multiple recommendations to create one single review for posting, and 4) not use recommendations at all. Reviewers are free to adapt any of the above methods or their combinations to write their review.

The task of developing review recommendation software will face the challenge of tracking how the reviewer used the recommendations. Combination of implicit and explicit evaluations of recommended reviews can be used to track how

reviewer used the recommendations. Current state-of-art indicates that there are mainly three different ways to get explicit feedback— (1) like/dislike, (2) ratings, and (3) text comments (Shapira *et al*., 2011). Review recommendation software can also use any of the above explicit feedback method to collect reviewers feedback. Additionally, implicit feedback technique for example button click, cursor hover, etc. may be used for keeping track of how many recommendations reviewers read through and which they liked. Further text match technique between the posted review and recommended review may be used to find which recommendations the reviewer used to post his or her review.

6.2.2 Enhancing reviewer-expert similarity

Reviewer-expert similarity enhancement is an interesting area for future work. Recall that in Section 5.3.1.1.1 in Chapter 5 we computed reviewer-expert similarity based on the evolution trend. The evolution was calculated on an annual basis in Section 5.1 in Chapter 5. In a future version, the evolution trend may be calculated on a monthly scale for finer estimation of evolution. This may result in more *precise* and *accurate* estimation of reviewer-reviewer similarity, which could improve review generation process. Another way to enhance reviewer-expert similarity is by incorporating more detailed reviewer and product profiles. Recall that in Chapter 4, the *reviewer profiles* were centered on the reviews they had posted such as review length; reviewing frequency; review helpfulness and so on. Adomavicius *et al*., (2005) have pointed that user profiles can include various user-specific characteristics such as age, gender, income, marital status, etc. The inclusion of these characteristics in the *reviewer profile* can produce effective recommendations for example the review posted by an expert with similar annual household income may be more effective to a reviewer than the review posted by

an expert with dissimilar income. Also, the inclusion of these characteristics can make reviewer modeling more meaningful leading to an effective recommendation generation.

Recall that the product feature in this thesis is limited to its categories such as Books, Baby, Grocery and gourmet foods, etc. Apart from this feature, product for example books can have other features such as price, title, genre, year of publication, author(s) and so on. *Product profile* can be created using aforementioned information and can be linked to the reviewer. Inclusion of product profiles would enhance reviewer-expert similarity calculation process and in turn enhance recommendation generation process. The review posted by an expert who has the same taste in books as the reviewer has a higher probability of being effective than the review posted by some other expert.

6.2.3 Diversifying recommendation generation

"Diversity is often a highly desirable feature in recommender systems" (Adomavicius *et al*., 2005). Diversification of recommendations through advanced filtering is an interesting future work. The past works of Zhang *et al*., (2010) and Adomavicius *et al*., (2005) can be starting point to implement this. To diversify recommendation, the reviews that are too *similar* to the reviews that the reviewer has already seen (in the recommendable list) should not be recommended, for example, different reviews describing the same feature of the product. Again the reviews that are too *different* must be filtered out from the recommendable list, as they may be describing an entirely different product (although in the same product category). In Books category, it may not be a good idea to recommend review describing science fiction novels to a reviewer who wants to review autobiography. The recommendable list must have a balance between reviews that are similar and yet different. For example, reviews that have diverse contents

and writing style, reviews that talk about multiple diverse features of the similar products, reviews that have both positive and negative feedbacks and so on. The reviewer should be able to see varieties in the recommended reviews.

References

Linden, Greg, Brent Smith, and Jeremy York. "Amazon.com recommendations: Item-to-item collaborative filtering." *Internet Computing, IEEE* 7.1 (2003): 76-80.

Bafna, Kushal, and Durga Toshniwal. "Feature based summarization of Customers' Reviews of Online products." *Procedia Computer Science* 22 (2013): 142-151.

McAuley, Julian, et al. "Image-based recommendations on styles and substitutes." *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015.

McAuley, Julian John, and Jure Leskovec. "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews."*Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013.

Odlyzko, Andrew M. "Internet traffic growth: Sources and implications."*ITCom 2003*. International Society for Optics and Photonics, 2003.

Pelleg, Dan, and Andrew W. Moore. "X-means: Extending K-means with Efficient Estimation of the Number of Clusters." *ICML*. Vol. 1. 2000.

Cai, Deng, Chiyuan Zhang, and Xiaofei He. "Unsupervised feature selection for multi-cluster data." *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.

Kumar, Varun, and Nisha Rathee. "Knowledge discovery from database using an integration of clustering and classification." *International Journal of Advanced Computer Science and Applications* 2.3 (2011): 29-33.

Zhao, Yongheng, and Yanxia Zhang. "Comparison of decision tree methods for finding active objects." *Advances in Space Research* 41.12 (2008): 1955-1959.

Zhang, Jason Q., Georgiana Craciun, and Dongwoo Shin. "When does electronic word-of-mouth matter? A study of consumer product reviews."*Journal of Business Research* 63.12 (2010): 1336-1341.

Ho, Jason YC, and Melanie Dempsey. "Viral marketing: Motivations to forward online content." *Journal of Business Research* 63.9 (2010): 1000-1006.

Juneja, Deepti, et al. "A novel approach to construct decision tree using quick C4. 5 algorithm." *Oriental Journal of Computer Science & Technology* 3.2 (2010): 305-310.

Witten, Ian H., and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

Witten, Ian H., et al. "Weka: Practical machine learning tools and techniques with Java implementations." (1999): 81-17.

Frias-Martinez, Enrique, Sherry Y. Chen, and Xiaohui Liu. "Survey of data mining approaches to user modeling for adaptive hypermedia." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 36.6 (2006): 734-749.

Quinlan, J. Ross. "C4. 5: Programming for machine learning." *Morgan Kauffmann* (1993): 38.

Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

Seo, Songwon. *A review and comparison of methods for detecting outliers in univariate data sets*. Diss. University of Pittsburgh, 2006.

Mudambi, Susan M., and David Schuff. "What makes a helpful review? A study of customer reviews on Amazon. com." *MIS quarterly* 34.1 (2010): 185-200.

Chen, Yubo, and Jinhong Xie. "Online consumer review: Word-of-mouth as a new element of marketing communication mix." *Management science* 54.3 (2008): 477-491.

Kumar, Nanda, and Izak Benbasat. "Research note: the influence of recommendations and consumer reviews on evaluations of websites."*Information Systems Research* 17.4 (2006): 425-439.

Kobsa, A. (2001). Generic user modeling systems. *User modeling and user-adapted interaction*, 11(1-2), 49-63.

Fischer, G. (2001). User modeling in human–computer interaction. *User modeling and user-adapted interaction*, 11(1-2), 65-86.

Romaine, Suzanne. *The language of children and adolescents: The acquisition of communicative competence*. Blackwell, 1984.

Puranam, Dinesh, Samuel Curtis Johnson, and Claire Cardie. "The Enrollment Effect: A Study of Amazon's Vine Program." *ACL 2014* (2014): 17.

Hu, Nan, Ling Liu, and Jie Jennifer Zhang. "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects." *Information Technology and Management* 9.3 (2008): 201-214.

Fink, Josef, Alfred Kobsa, and Andreas Nill. "Adaptable and adaptive information provision for all users, including disabled and elderly people."*New review of Hypermedia and Multimedia* 4.1 (1998): 163-188.

Jameson, Anthony. "Adaptive interfaces and agents." *Human-Computer Interaction: Design Issues, Solutions, and Applications* 105 (2009): 105-130.

Litman, Diane J., and Shimei Pan. "Designing and evaluating an adaptive spoken dialogue system." *User Modeling and User-Adapted Interaction* 12.2-3 (2002): 111-137.

Komatani, Kazunori, et al. "User modeling in spoken dialogue systems to generate flexible guidance." *User Modeling and User-Adapted Interaction*15.1-2 (2005): 169-183.

Suraweera, Pramuditha, and Antonija Mitrovic. "An intelligent tutoring system for entity relationship modelling." *International Journal of Artificial Intelligence in Education* 14.3, 4 (2004): 375-417.

Gutkauf, Bernd, Stefanie Thies, and Gitta Domik. "A user-adaptive chart editing system based on user modeling and critiquing." *User Modeling*. Springer Vienna, 1997.

Dalamagas, Theodore, et al. "Mining user navigation patterns for personalizing topic directories." *Proceedings of the 9th annual ACM international workshop on Web information and data management*. ACM, 2007.

Porter, Joshua. *Designing for the Social Web, eBook*. Peachpit Press, 2010.

Gervasio, Melinda T., et al. "Active preference learning for personalized calendar scheduling assistance." *Proceedings of the 10th international conference on Intelligent user interfaces*. ACM, 2005.

Zhou, Michelle X., and Vikram Aggarwal. "An optimization-based approach to dynamic data content selection in intelligent multimedia interfaces."*Proceedings of the 17th annual ACM symposium on User interface software and technology*. ACM, 2004.

Hu, Nan, Ling Liu, and Jie Jennifer Zhang. "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects." *Information Technology and Management* 9.3 (2008): 201-214.

Chevalier, Judith A., and Dina Mayzlin. "The effect of word of mouth on sales: Online book reviews." *Journal of marketing research* 43.3 (2006): 345-354.

Lee, Jumin, Do-Hyung Park, and Ingoo Han. "The effect of negative online consumer reviews on product attitude: An information processing view." *Electronic commerce research and applications* 7.3 (2008): 341-352.

Park, Do-Hyung, and Jumin Lee. "eWOM overload and its effect on consumer behavioral intention depending on consumer involvement." *Electronic Commerce Research and Applications* 7.4 (2009): 386-398.

Dellarocas, Chrysanthos, Guodong Gao, and Ritu Narayan. "Are consumers more likely to contribute online reviews for hit or niche products?." *Journal of Management Information Systems* 27.2 (2010): 127-158.

Pan, Yue, and Jason Q. Zhang. "Born unequal: a study of the helpfulness of user-generated product reviews." *Journal of Retailing* 87.4 (2011): 598-612.

Nelson, Phillip. "Information and consumer behavior." *Journal of political economy* 78.2 (1970): 311-329.

Korfiatis, Nikolaos, Elena García-Bariocanal, and Salvador Sánchez-Alonso. "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content." *Electronic Commerce Research and Applications* 11.3 (2012): 205-217.

Ricci, Francesco, Lior Rokach, and Bracha Shapira. *Introduction to recommender systems handbook*. Springer US, 2011.

Pham, Xuan Hau, et al. "-Spear: A New Method for Expert Based Recommendation Systems." *Cybernetics and Systems* 45.2 (2014): 165-179.

Herlocker, Jonathan L., et al. "Evaluating collaborative filtering recommender systems." *ACM Transactions on Information Systems (TOIS)* 22.1 (2004): 5-53.

Paris, Cecile. *User modelling in text generation*. Bloomsbury Publishing, 2015.

Anzai, Yuichiro. *Pattern Recognition & Machine Learning*. Elsevier, 2012.

Maglogiannis, Ilias G. *Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*. Vol. 160. Ios Press, 2007.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*. Vol. 20. Siam, 2007.

Strehl, Alexander, Joydeep Ghosh, and Raymond Mooney. "Impact of similarity measures on web-page clustering." *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*. 2000.

Clatworthy, Jane, et al. "The use and reporting of cluster analysis in health psychology: A review." *British journal of health psychology* 10.3 (2005): 329-358.

Wedel, Michel, and Wagner A. Kamakura. *Market segmentation: Conceptual and methodological foundations*. Vol. 8. Springer Science & Business Media, 2012.

Jain, Anil K. "Data clustering: 50 years beyond K-means." *Pattern recognition letters* 31.8 (2010): 651-666.

Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "*Supervised machine learning: A review of classification techniques*." (2007): 3-24.

Beck, Joseph E., et al. "*Predicting student help-request behavior in an intelligent tutor for reading*." *International Conference on User Modeling*. Springer Berlin Heidelberg, 2003.

Go, Alec, Richa Bhayani, and Lei Huang. "*Twitter sentiment classification using distant supervision*." *CS224N Project Report, Stanford* 1 (2009): 12.

Pang, Bo, and Lillian Lee. "*Opinion mining and sentiment analysis*." *Foundations and trends in information retrieval* 2.1-2 (2008): 1-135.

Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5.1 (2012): 1-167.

McGlohon, Mary, Natalie S. Glance, and Zach Reiter. "*Star Quality: Aggregating Reviews to Rank Products and Merchants*." *ICWSM*. 2010.

Hai, Zhen, Kuiyu Chang, and Jung-jae Kim. "*Implicit feature identification via co-occurrence association rule mining*." *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer Berlin Heidelberg, 2011.

Zhang, Wenbin, and Steven Skiena. "*Trading Strategies to Exploit Blog and News Sentiment*." *ICWSM*. 2010.

Madhoushi, Zohreh, Abdul Razak Hamdan, and Suhaila Zainudin. "Sentiment analysis techniques in recent works." *Science and Information Conference (SAI), 2015*. IEEE, 2015.

Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." *Eighth International AAAI Conference on Weblogs and Social Media*. 2014.

Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Vol. 1631. 2013.

Ribeiro, Filipe Nunes, et al. "A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods." *arXiv preprint arXiv:1512.01818* (2015).

Otterbacher, Jahna. "'Helpfulness' in online communities: a measure of message quality." *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2009.

McAuley, Julian, and Alex Yang. "Addressing Complex and Subjective Product-Related Queries with Customer Reviews." *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016.

Bouckaert, Remco R., et al. "WEKA Manual for Version 3-7-8." *Hamilton, New Zealand* (2013).

Draper, N., and H. Smith. "Applied regression analysis: Wiley Interscience." *New York* (1998): 505-553.

Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." (1973): 32-57.

Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." *IEEE transactions on knowledge and data engineering*17.6 (2005): 734-749.

Massa, Paolo, and Paolo Avesani. "Trust-aware collaborative filtering for recommender systems." *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer Berlin Heidelberg, 2004.

Shapira, Bracha, et al. "Recommender Systems Handbook." (2011).

Chapter 7
Appendices

There are seven appendices. In Appendix A Amazon review data is presented to highlight the before and after of data cleaning process. The data cleaning process was carried to remove data imbalance as well as inactive reviewers. Appendix B shows detail statistics of each clusters after *X*-means clustering was performed. Appendices C and D contain J48 decision trees and confusion matrices after performing 10-fold cross validation respectively. Appendix E shows how each reviewer class evolves over time. Appendix F shows the results of sentiment analysis i.e., correlation between helpfulness and review tone; and overall and review tone. Finally, Appendix G shows the data structure of review and reviewer class.

A. Data cleaning

The trend of data for seven products categories is plotted to see how they change over time. Graphs contain review count; user count and product count in log with respect to time in month. If there is any data imbalance (such as abrupt change), the data is cleaned to address data imbalance.

Electronics

Amazon electronics review data contains reviews from 1998 to 2014. Below is a graph of review count; user count and product count in log with respect to time in month.

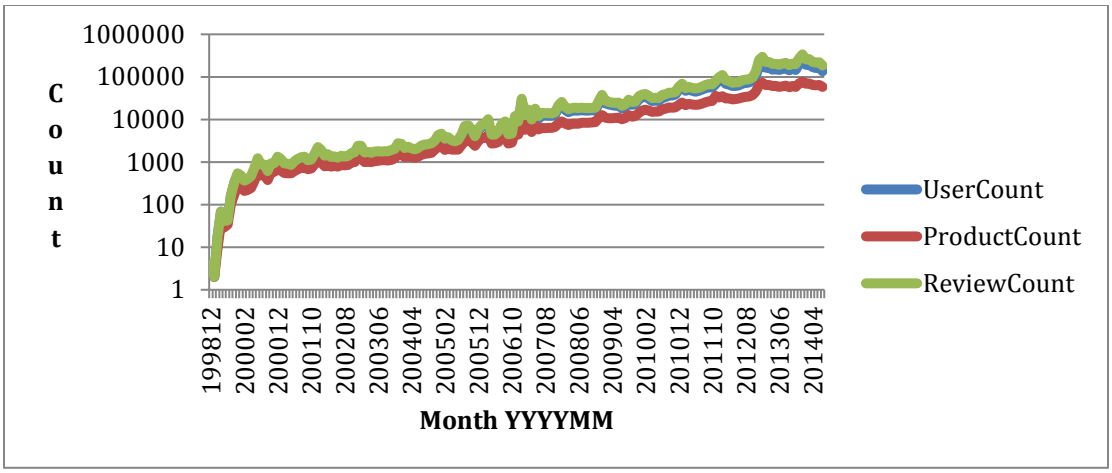

Figure 7.1: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Electronics" category

As we can see in the above graph, the data is imbalanced, growing exponentially before year 2000. However, the growth from 2003 to 2013 is more consistent and gradually increasing. So we choose to remove data that was exponentially increasing and keep balanced data from year 2003 to 2013. Below is the graph that shows review count; user count and product count in log with respect to time in month from year 2003 to 2013.

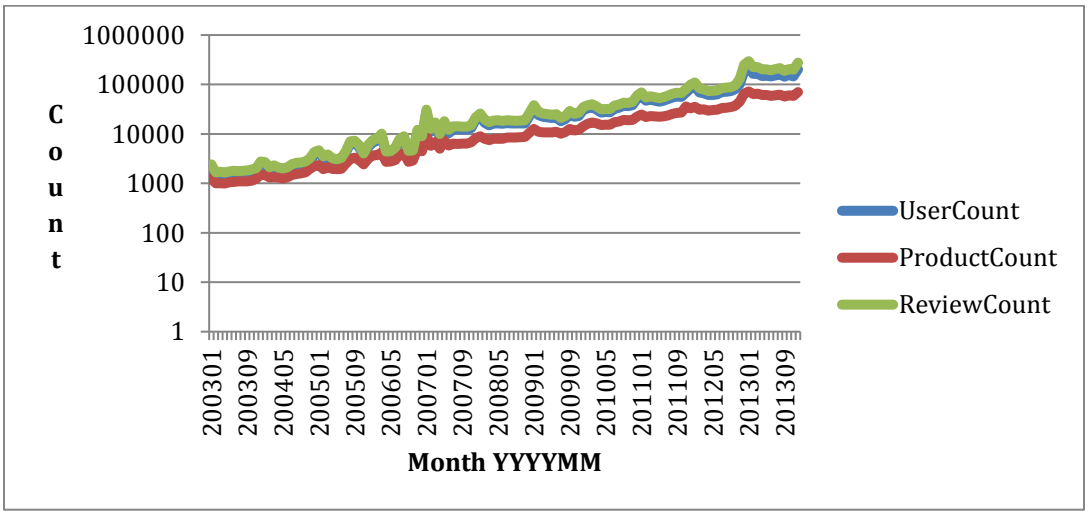

Figure 7.2: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) after cleaning the data in "Electronics" category

After cleaning, amazon electronics review data from Jan 2003 to Dec 2013 looks more gradual.

Cellphones and accessories

Amazon cellphones and accessories review data contains reviews from 1991 to 2014. Below is a graph of review count; user count and product count in log with respect to time in month.

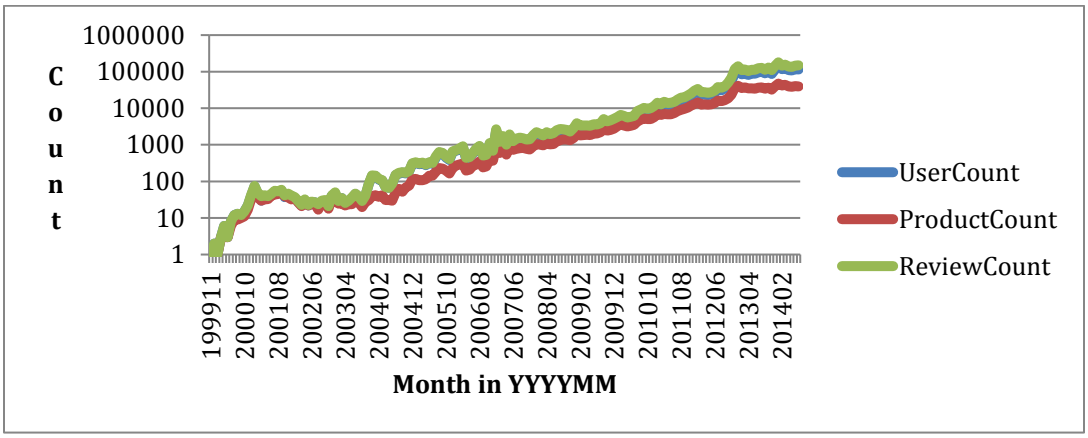

Figure 7.3: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Cellphones & accessories" category

As we can see in the above graph, the data is imbalanced, growing exponentially before year 2000. However, the growth from 2003 to 2013 is more consistent and gradually increasing. So we choose to remove data that was exponentially increasing and keep balanced data from year 2003 to 2013. Below is the graph that shows review count; user count and product count in log with respect to time in month from year 2003 to 2013.

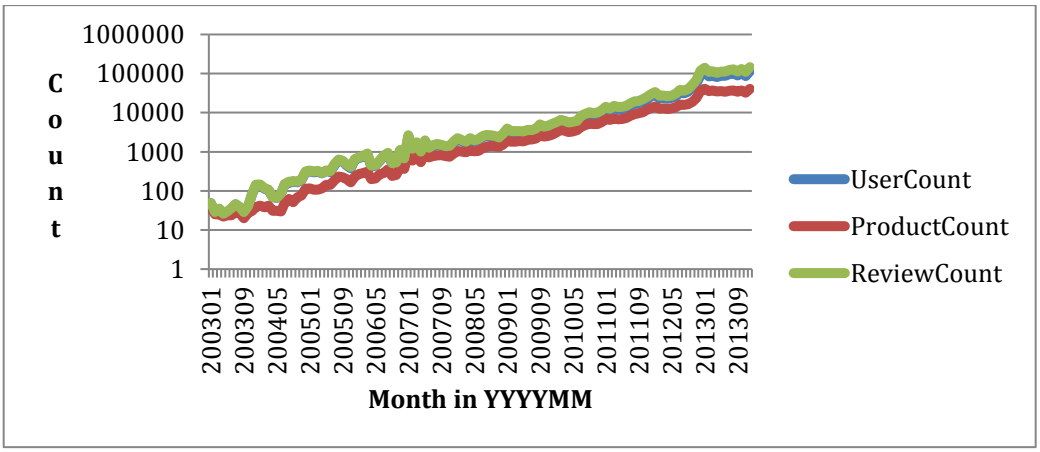

Figure 7.4: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) after cleaning the data in "Cellphones & accessories" category

After cleaning, amazon cellphones and accessories review data from Jan 2003 to Dec 2013 looks more gradual.

Grocery and gourmet food

Amazon grocery and gourmet food review data contains reviews from 2000 to 2014. Below is a graph of review count; user count and product count in log with respect to time in month.

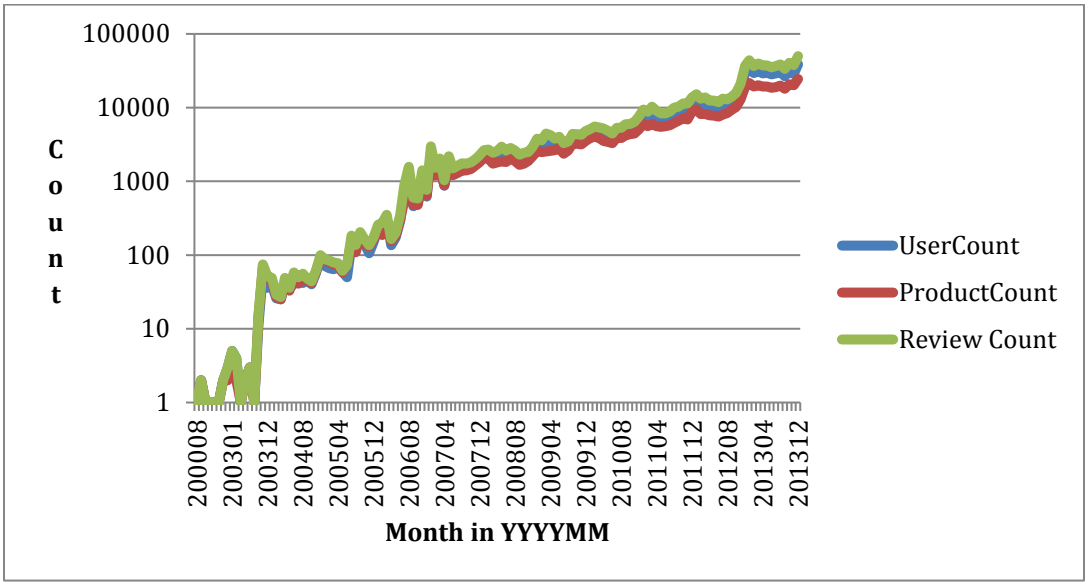

Figure 7.5: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Grocery & gourmet food" category

As we can see in the above graph, the data is imbalanced, growing exponentially before year 2000. However, the growth from 2004 to 2013 is more consistent and gradually increasing. So we choose to remove data that was exponentially increasing and keep balanced data from year 2004 to 2013. Below is the graph that shows review count; user count and product count in log with respect to time in month from year 2004 to 2013.

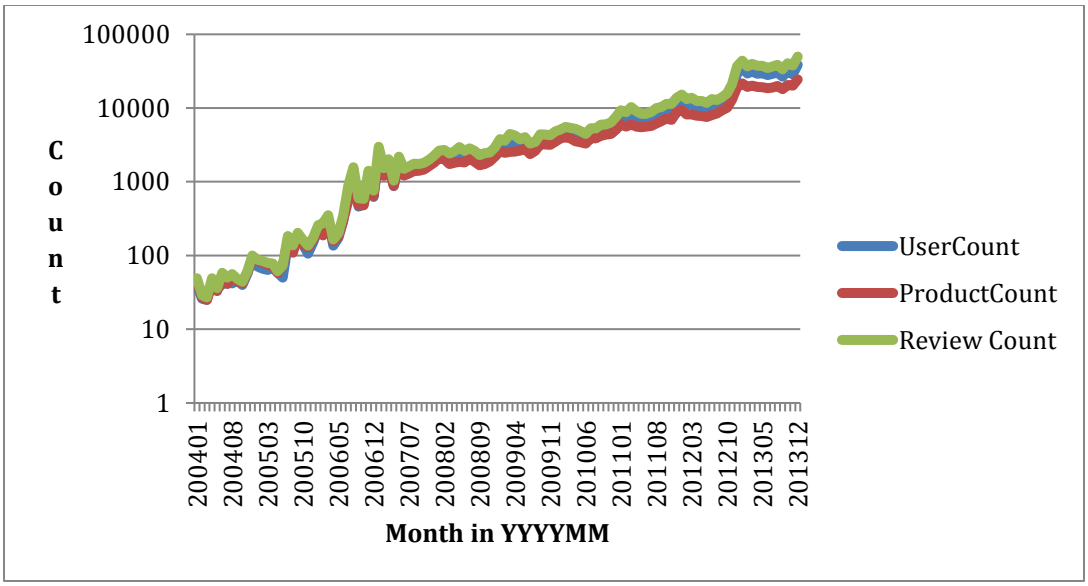

Figure 7.6: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) after cleaning the data in "Grocery & gourmet food" category

After cleaning, amazon grocery and gourmet food data from Jan 2004 to Dec 2013 looks more gradual.

Health and personal care

Amazon health and personal care review data contains reviews from 1998 to 2014. Below is a graph of review count; user count and product count in log with respect to time in month.

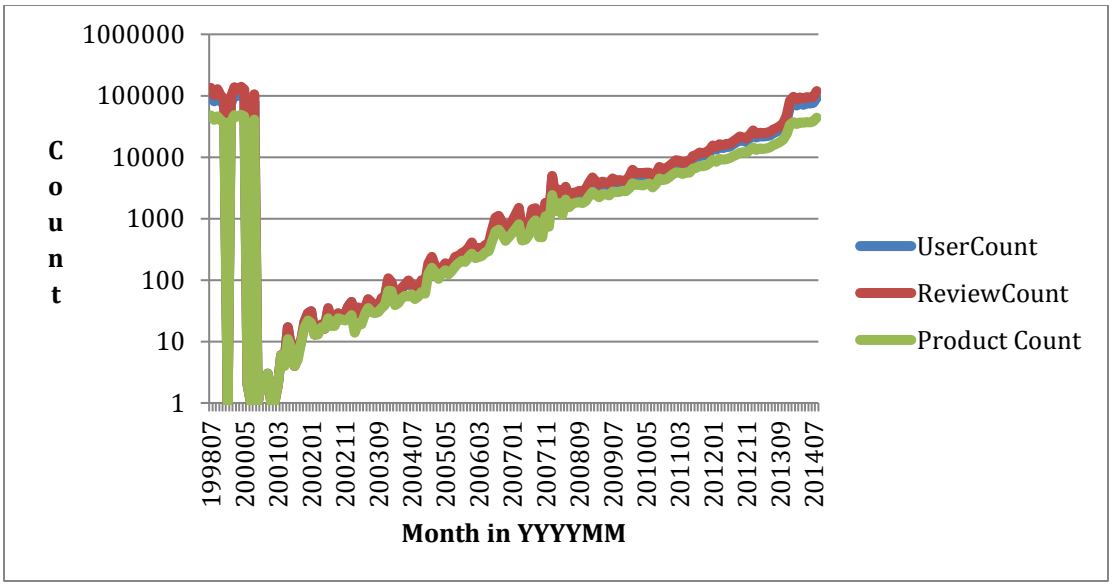

Figure 7.7: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Health & personal care" category

As we can see in the above graph, the data is imbalanced, growing exponentially before year 2000. However, the growth from 2003 to 2013 is more consistent and gradually increasing. So we choose to remove data that was exponentially increasing and keep balanced data from year 2003 to 2013. Below is the graph that shows review count; user count and product count in log with respect to time in month from year 2003 to 2013.

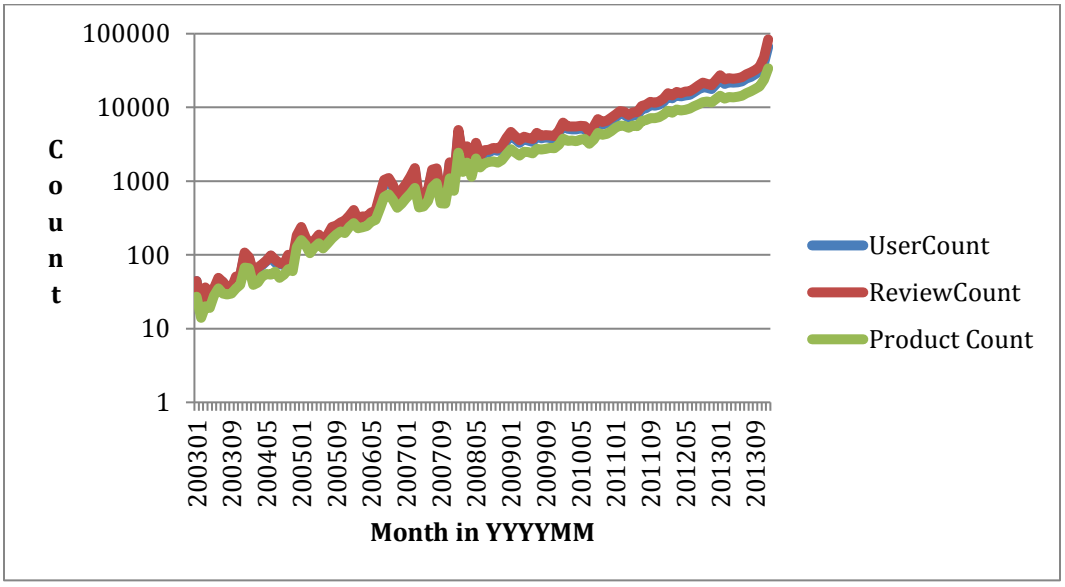

Figure 7.8: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) after cleaning the data in "Health & personal care" category

After cleaning, amazon health and personal data from Jan 2003 to Dec 2013 looks more gradual. We will repeat this process for amazon office products review data.

Office products

Amazon office products review data contains reviews from 1998 to 2014. Below is a graph of review count; user count and product count in log with respect to time in month.

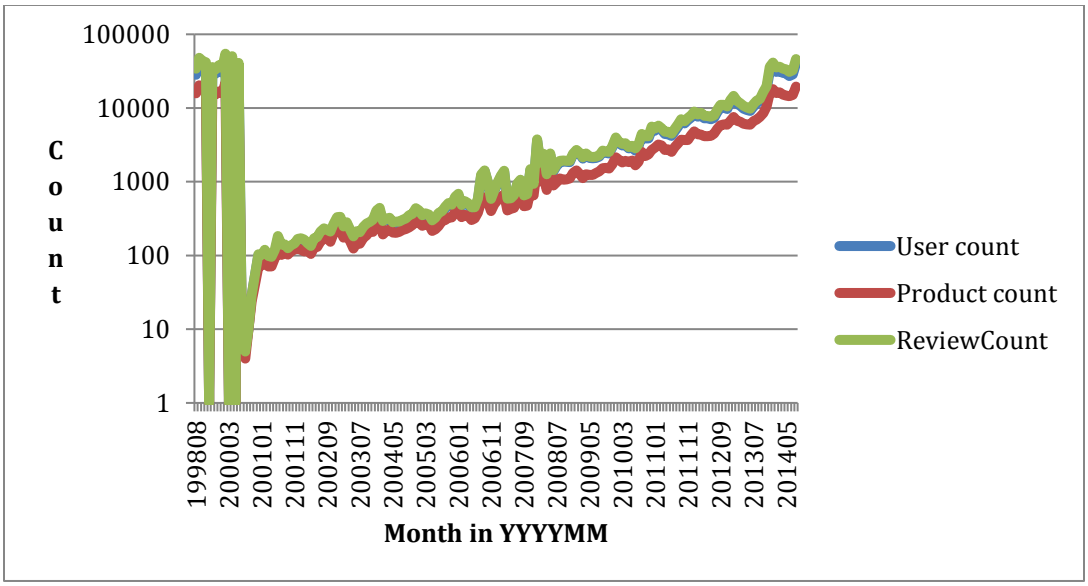

Figure 7.9: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Office product" category

As we can see in the above graph, the data is imbalanced, growing exponentially before year 2000. However, the growth from 2002 to 2013 is more consistent and gradually increasing. So we choose to remove data that was exponentially increasing and keep balanced data from year 2002 to 2013. Below is the graph that shows review count; user count and product count in log with respect to time in month from year 2002 to 2013.

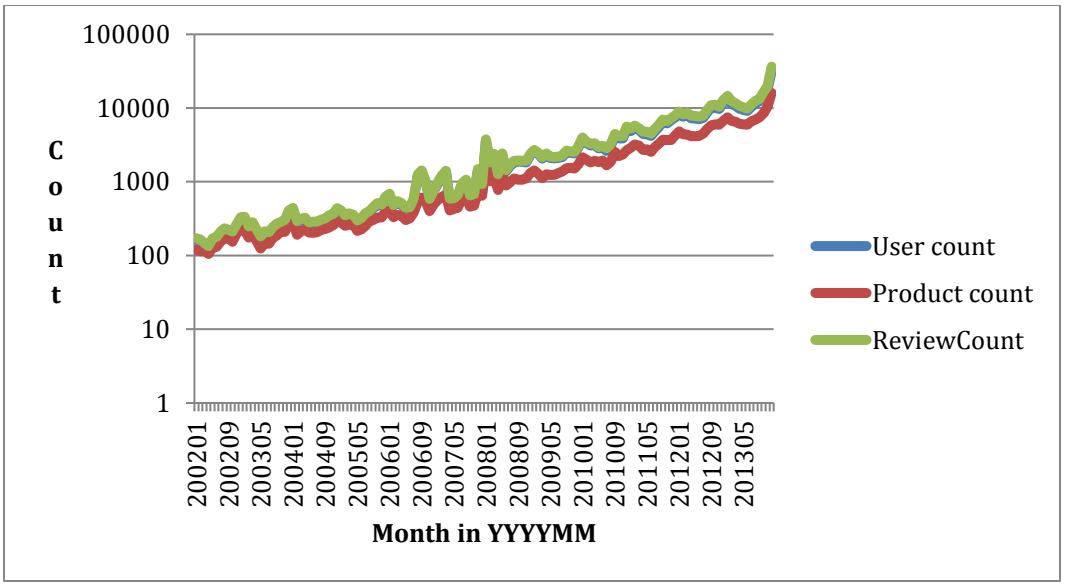

Figure 7.10: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) after cleaning the data in "Office product" category

After cleaning, amazon office product data from Jan 2002 to Dec 2013 looks more gradual as seen in above graph.

Baby

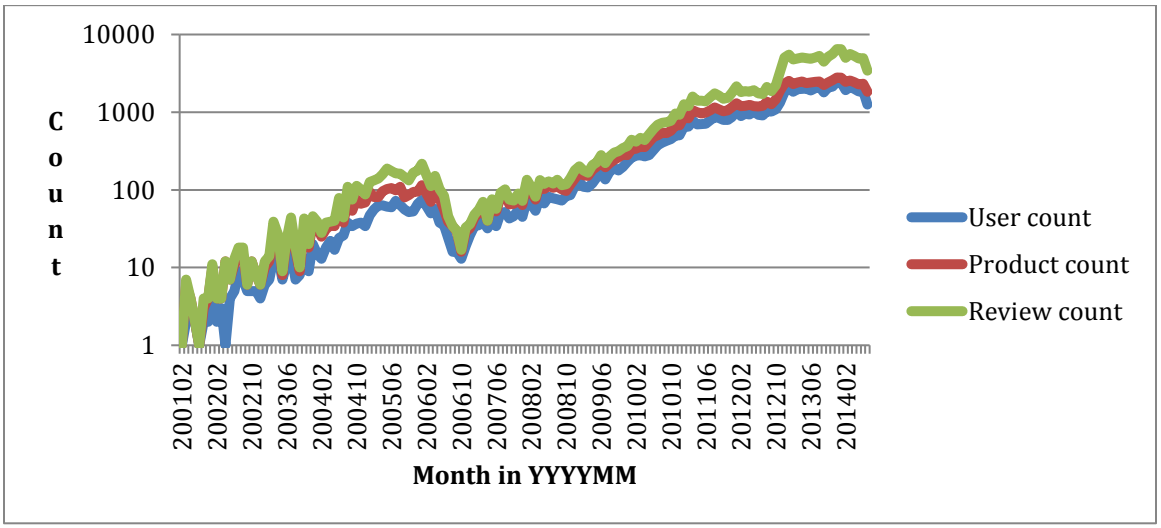

Figure 7.11: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Baby" category

As we can see in the above graph, the data is imbalanced, growing exponentially before year 2000. However, the growth from 2002 to 2013 is more consistent and gradually

increasing. So we choose to remove data that was exponentially increasing and keep balanced data from year 2002 to 2013.

Below is the graph that shows review count; user count and product count in log with respect to time in month from year 2002 to 2013.

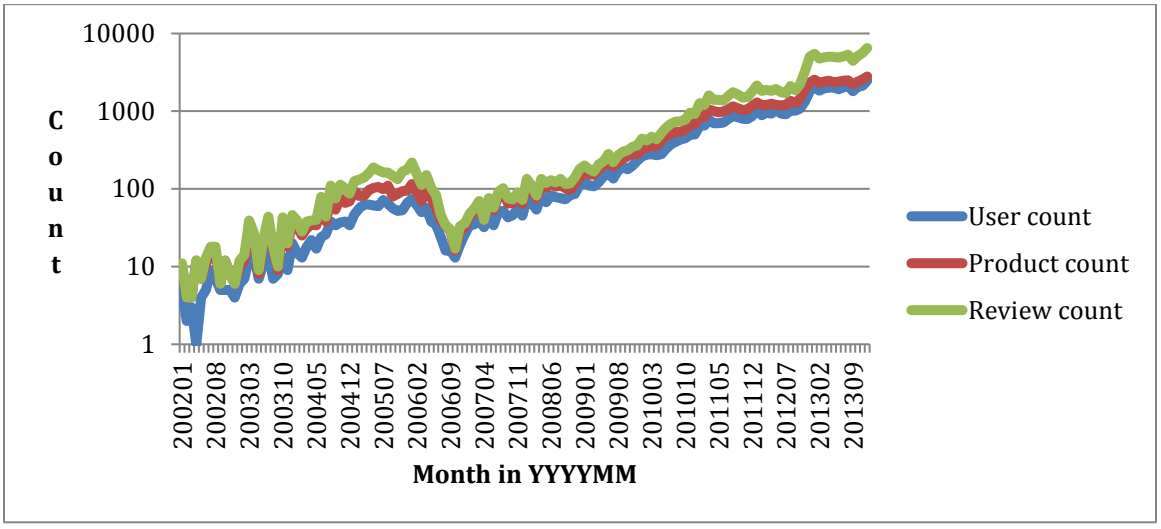

Figure 7.12: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) after cleaning the data in "Baby" category

After cleaning, amazon baby product data from Jan 2002 to Dec 2013 looks more gradual as seen in above graph.

Beauty

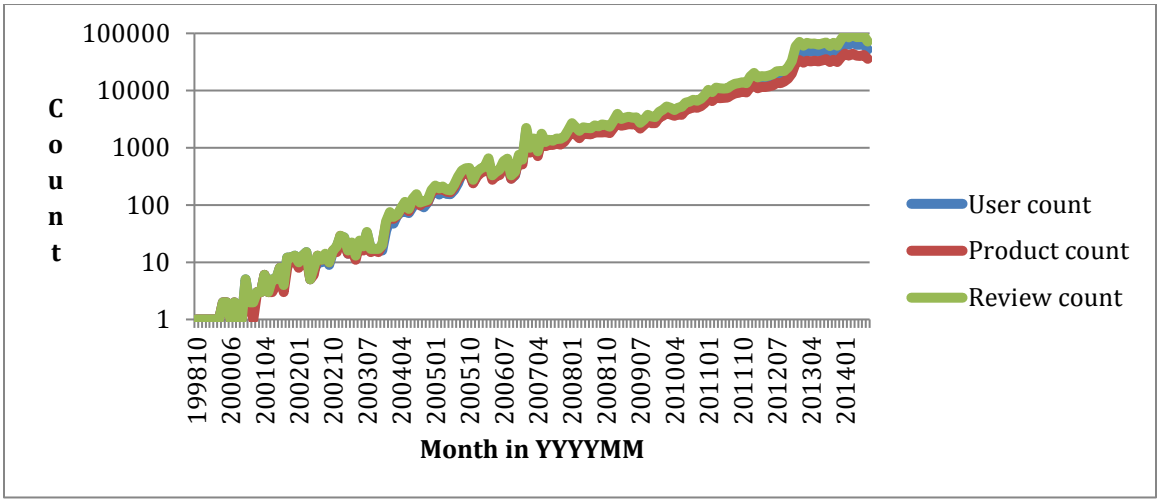

Figure 7.13: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Beauty" category

As we can see in the above graph, the data is imbalanced, growing exponentially before year 2000. However, the growth from 2002 to 2013 is more consistent and gradually increasing. So we choose to remove data that was exponentially increasing and keep balanced data from year 2002 to 2013. Below is the graph that shows review count; user count and product count in log with respect to time in month from year 2002 to 2013.

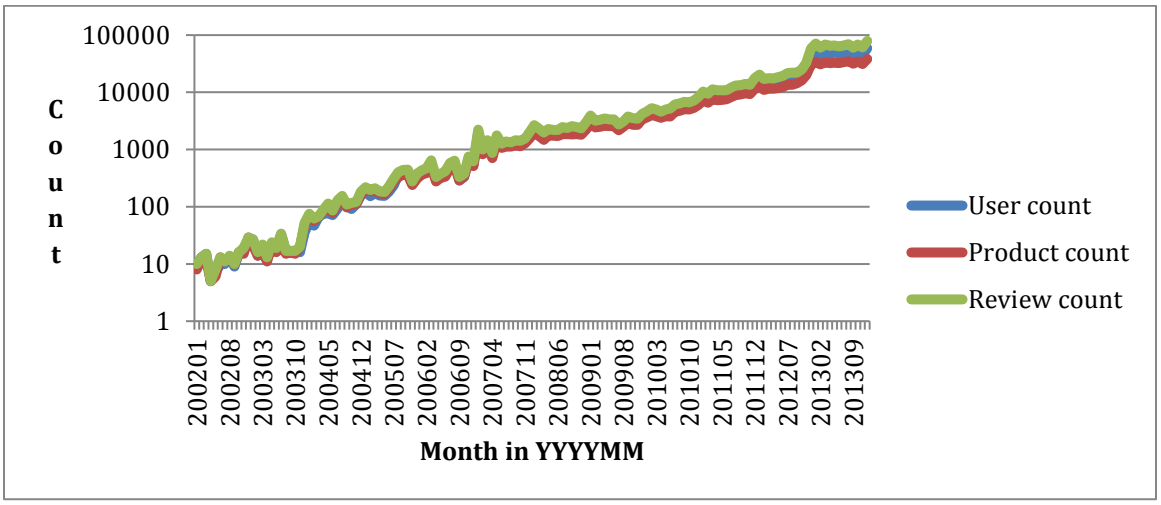

Figure 7.14: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Beauty" category

After cleaning, amazon beauty product data from Jan 2002 to Dec 2013 looks more gradual as seen in above graph.

Pet supplies

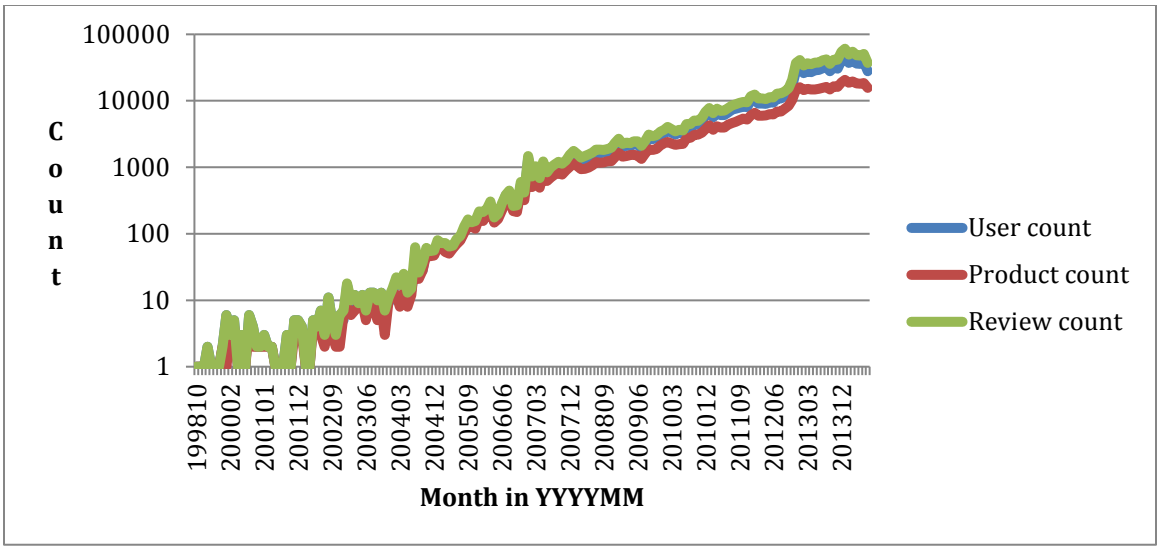

Figure 7.15: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) before cleaning the data in "Pet supplies" category

As we can see in the above graph, the data is unbalanced, growing exponentially before year 2000. However, the growth from 2002 to 2013 is more consistent and gradually increasing. So we choose to remove data that was exponentially increasing and keep balanced data from year 2002 to 2013. Below is the graph that shows review count; user count and product count in log with respect to time in month from year 2002 to 2013.

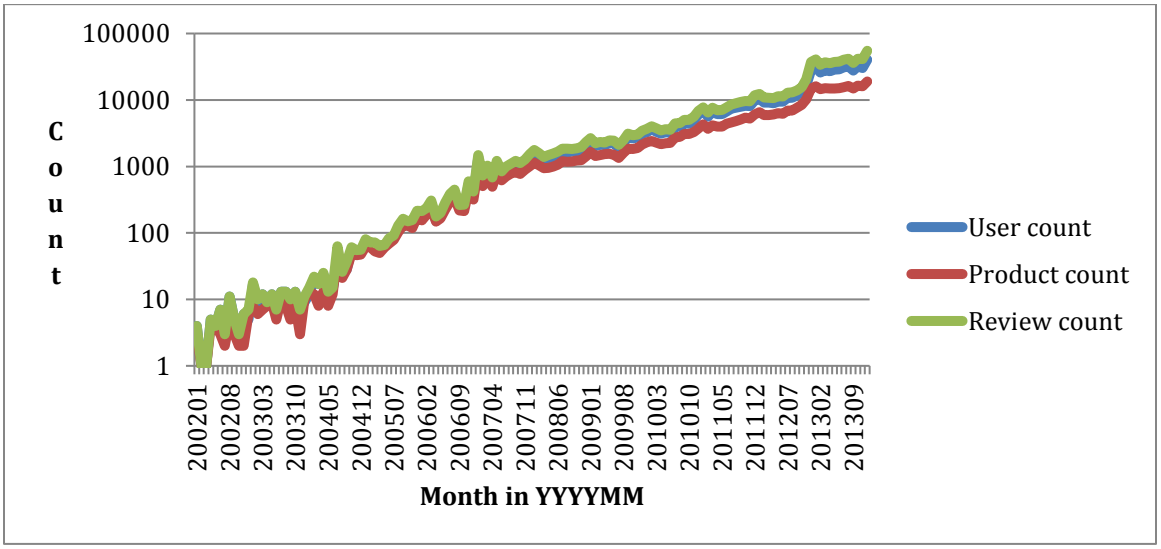

Figure 7.16: Graph showing the distribution of Review Count, User Count and Product Count (in log scales) with respect to Date (YYYYMM) after cleaning the data in "Pet supplies" category

After cleaning, amazon pet supplies data from Jan 2002 to Dec 2013 looks more gradual as seen in above graph.

Data cleaning to remove large proportion of inactive users

There is a large number of users who are active for very few months that makes it harder to find their characteristics to understand their reviewing pattern. To remove these inactive users from consideration, we identify the threshold value on number of months that differentiates between active and inactive users. We observe the number of users active over time and how the user count changes in all seven-product categories and then remove inactive users.

Electronics

Below is the table that contains top 15 values of active month count, user count and the slope of user count. The slope indicates the rate at which user count is increasing or decreasing with respect to month count for amazon electronics review data.

Month Count	User Count	Slope
1	2639358	13340
2	444015	2962.1
3	144304	1196.2
4	62058	**614.75**
5	31363	361.05
6	17384	230.7
7	10510	157.49
8	6662	112.49
9	4518	83.63
10	3119	63.6
11	2197	49.57
12	1635	39.63
13	1185	32.03
14	913	26.64
15	697	22.45

Table 7.1: Distribution of User count, with respect to their active month (i.e. number of month they have posted a review) in "Electronics" category

From the values of the slope in Table 2, we can observe that the graph descends abruptly till active month is 4 and descends gradually after that i.e. the number of users who are

active for 4 months or more are more-or-less linear with respect to active months. Also from the table 1, we see that the number of users who are active for 3 months or less grow (or shrink) almost exponentially.

Below is the graph of user count in log with respect to month count before and after removing users active for 4 month:

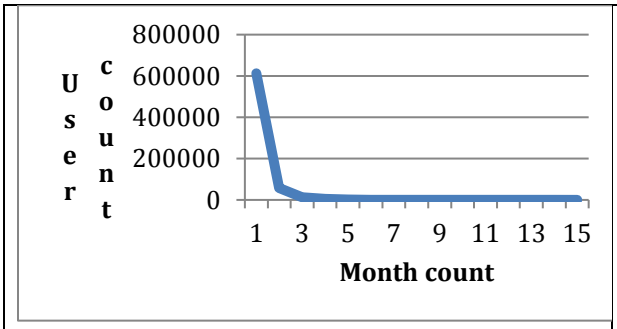

Figure 7.17: User count vs. Month count before removing users active for less than 4 month in "Electronics" category

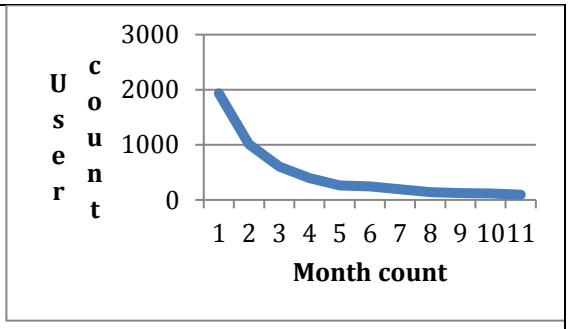

Figure 7.18: User count vs. Month count after removing users active for less than 4 month in "Electronics" category

Looking at the above graphs we can say that the active user count is balanced after removing the users active for less than 4 month. We will repeat this process for amazon cellphones and accessories review data.

Cellphones and accessories

Below is the table that contains top 15 values of active month count, user count and the slope of user count. The slope indicates the rate at which user count is increasing or decreasing with respect to month count for amazon cellphones and accessories review data.

Month Count	User Count	Slope
1	1916772	34790
2	246238	5678
3	60733	1733.2
4	20285	**705.46**
5	8144	343.46
6	3867	191.39
7	1969	114.57
8	1123	73.413
9	622	47.46

10	366	32.72
11	233	24.3
12	180	19.4
13	112	13.71
14	84	10.91
15	62	8.2

Table 7.2: Distribution of User count, with respect to their active month (i.e. number of month they have posted a review) in "Cellphones & accessories" category

From the values of the slope in table 3, we can observe that the graph descends abruptly till active month is 4 and descends gradually after that i.e. the number of users who are active for 4 months or more are more-or-less linear with respect to active months. Also from the table 1, we see that the number of users who are active for 3 months or less grow (or shrink) almost exponentially.

Below is the graph of user count with respect to month count before and after removing users active for 4 month:

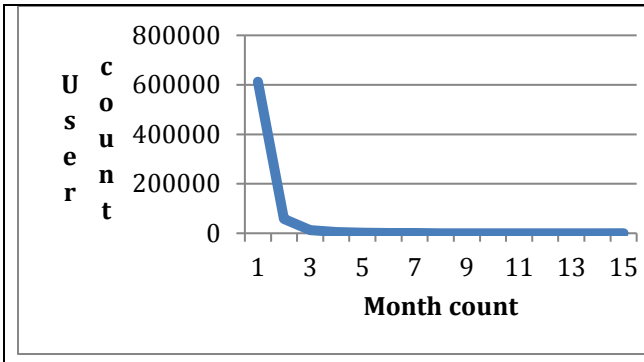

Figure 7.19: User count vs. Month count before removing users active for less than 4 month in "Cellphones & accessories" category

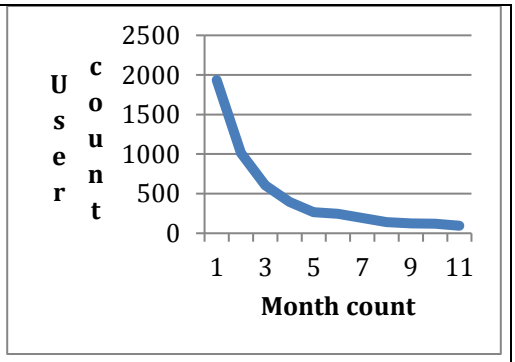

Figure 7.20: User count vs. Month count after removing users active for less than 4 month in "Cellphones & accessories" category

Looking at the above graphs we can say that the active user count is balanced after removing the users active for less than 4 month. We will repeat this process for amazon Grocery and gourmet food review data.

Grocery and gourmet food

Below is the table that contains top 15 values of active month count, user count and the slope of user count. The slope indicates the rate at which user count is increasing or

decreasing with respect to month count for amazon grocery and gourmet food review data.

Month Count	User Count	Slope
1	486220	4836.5
2	60398	802.35
3	17071	**291.74**
4	6710	146.13
5	3467	89.67
6	2040	60.52
7	1264	43.28
8	858	32.84
9	591	25.84
10	445	21.37
11	353	18.12
12	273	15.5
13	220	13.71
14	204	12.46
15	198	11.6

Table 7.3: Distribution of User count, with respect to their active month (i.e. number of month they have posted a review) in "Grocery & gourmet food" category

From the values of the slope in table 4, we can observe that the graph descends abruptly till active month is 3 and descends gradually after that i.e. the number of users who are active for 3 months or more are more-or-less linear with respect to active months. Also from the Table 3, we see that the number of users who are active for 2 months or less grow (or shrink) almost exponentially. Below is the graph of user count with respect to month count before and after removing users active for 3 month:

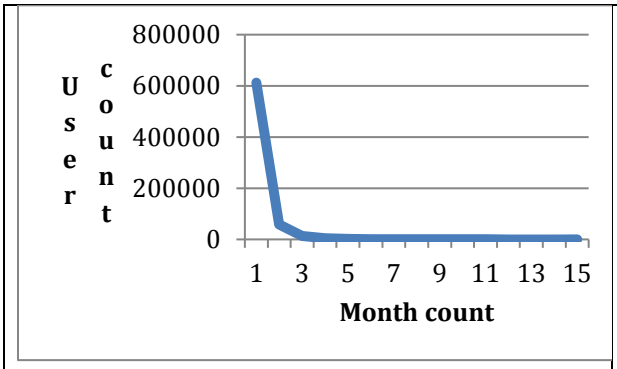

Figure 7.21: User count vs. Month count before removing users active for less than 3 month in "Grocery & gourmet food" category

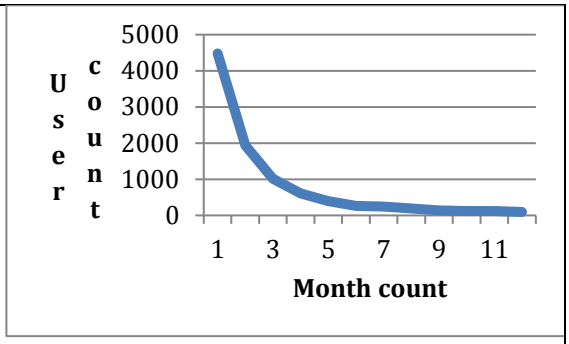

Figure 7.22: User count vs. Month count after removing users active for less than 3 month in "Grocery & gourmet food" category

Looking at the above graphs we can say that the active user count is balanced after removing the users active for less than 4 month. We will repeat this process for amazon health and personal care review data.

Health and personal care

Below is the table that contains top 15 values of active month count, user count and the slope of user count. The slope indicates the rate at which user count is increasing or decreasing with respect to month count for amazon health and personal care review data.

Month count	User count	Slope
1	1138669	4878.1
2	143619	790.1
3	38516	**264.72**
4	14555	121.66
5	6722	66.9
6	3470	41.37
7	2093	28.23
8	1315	20.21
9	810	15.17
10	610	12.26
11	480	10.022
12	337	8.16
13	234	6.95
14	210	6.28
15	192	5.59

Table 7.4: Distribution of User count, with respect to their active month (i.e. number of month they have posted a review) in "Health & personal care" category

From the values of the slope in Table 5, we can observe that the graph descends abruptly till active month is 3 and descends gradually after that i.e. the number of users who are active for 3 months or more are more-or-less linear with respect to active months. Also from the table 3, we see that the number of users who are active for 2 months or less grow (or shrink) almost exponentially.

Below is the graph of user count with respect to month count before and after removing users active for 3 month:

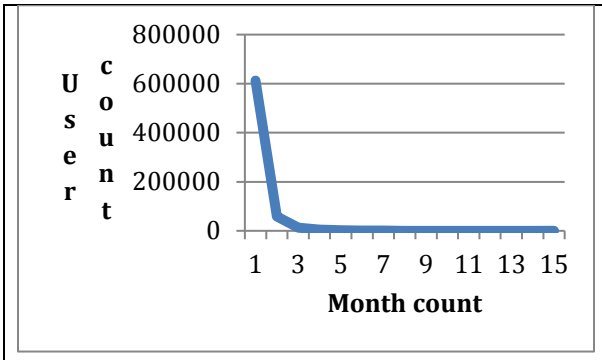

Figure 7.23: User count vs. Month count before removing users active for less than 3 month in "Health & personal care" category

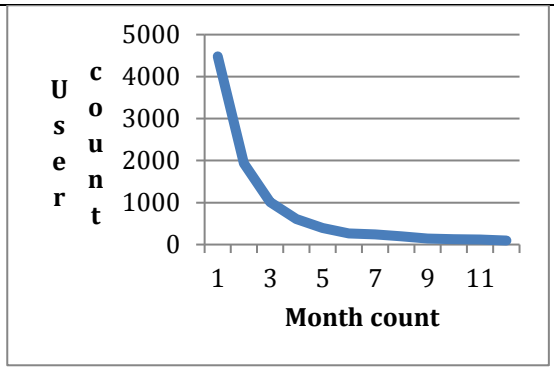

Figure 7.24: User count vs. Month count after removing users active for less than 3 month in "Health & personal care" category

Looking at the above graphs we can say that the active user count is balanced after removing the users active for less than 3 month. We will repeat this process for amazon office products review data.

Office product

Below is the table that contains top 15 values of active month count, user count and the slope of user count. The slope indicates the rate at which user count is increasing or decreasing with respect to month count for amazon office products review data.

Month count	User count	Slope
1	612503	4415.9
2	58353	534.05
3	13044	**152.55**
4	4481	65.7

5	1935	35.65
6	1013	22.89
7	604	16.44
8	397	12.8
9	265	10.6
10	245	9.44
11	195	8.06
12	141	6.94
13	124	6.32
14	120	5.69
15	95	4.74

Table 7.5: Distribution of User count, with respect to their active month (i.e. number of month they have posted a review) in "Office product" category

From the values of the slope in Table 6, we can observe that the graph descends abruptly till active month is 3 and descends gradually after that i.e. the number of users who are active for 3 months or more are more-or-less linear with respect to active months. Also from the table 3, we see that the number of users who are active for 2 months or less grow (or shrink) almost exponentially.

Below is the graph of user count with respect to month count before and after removing users active for 3 month:

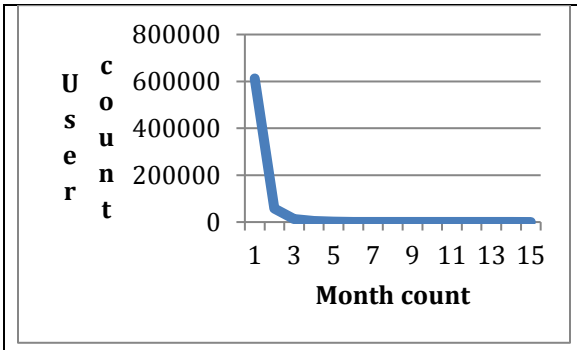

Figure 7.25: User count vs. Month count before removing users active for less than 3 month in "Office product" category

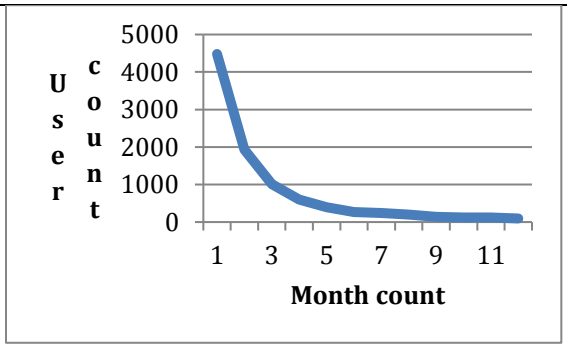

Figure 7.26: User count vs. Month count after removing users active for less than 3 month in "Office product" category

Looking at the above graphs we can say that the active user count is balanced after removing the users active for less than 3 month.

Beauty

Below is the table that contains top 15 values of active month count, user count and the slope of user count. The slope indicates the rate at which user count is increasing or decreasing with respect to month count for amazon beauty review data.

Month count	User count	Slope
1	749246	21194
2	87353	3316.2
3	23443	**1142**
4	8787	536.12
5	4157	303.81
6	2190	189.53
7	1240	128.18
8	811	95.702
9	557	73.714
10	384	57
11	268	45
12	198	38.5
13	156	37.5
14	122	41
15	81	

Table 7.6: Distribution of User count, with respect to their active month (i.e. number of month they have posted a review) in "Beauty" category

Below is the graph of user count with respect to month count before and after removing users active for 3 month:

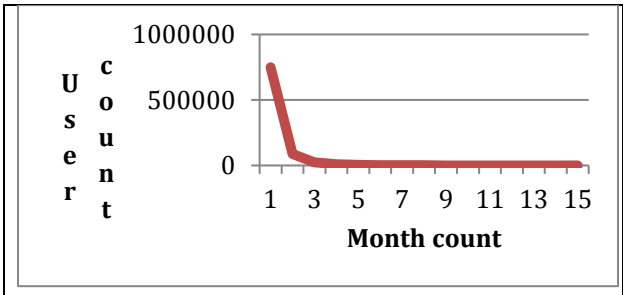

Figure 7.27: User count vs. Month count before removing users active for less than 3 month in "Beauty" category

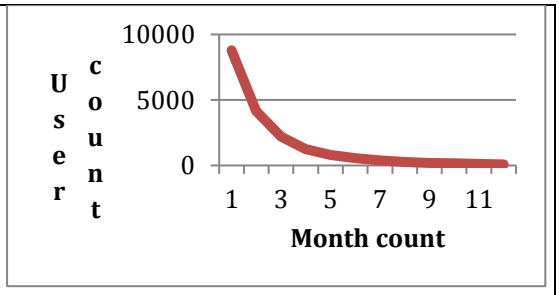

Figure 7.28: User count vs. Month count after removing users active for less than 3 month in "Beauty" category

Pet Supplies

Below is the table that contains top 15 values of active month count, user count and the slope of user count. The slope indicates the rate at which user count is increasing or decreasing with respect to month count for amazon beauty review data.

Month count	User count	Slope
1	466973	13383
2	60079	2299.7
3	16428	**810.16**
4	6379	387.72
5	3030	215.76
6	1406	126.58
7	834	87.867
8	557	63.69
9	303	42.357
10	205	33.486
11	143	28
12	112	25.8
13	73	20
14	52	19
15	33	

Table 7.7: Distribution of User count, with respect to their active month (i.e. number of month they have posted a review) in "Pet supplies" category

Below is the graph of user count with respect to month count before and after removing users active for 3 month:

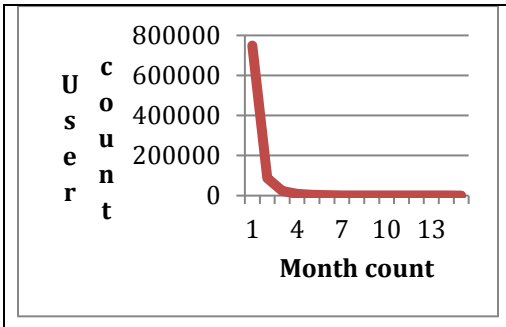

Figure 7.29: User count vs. Month count before removing users active for less than 3 month in "Pet supplies" category

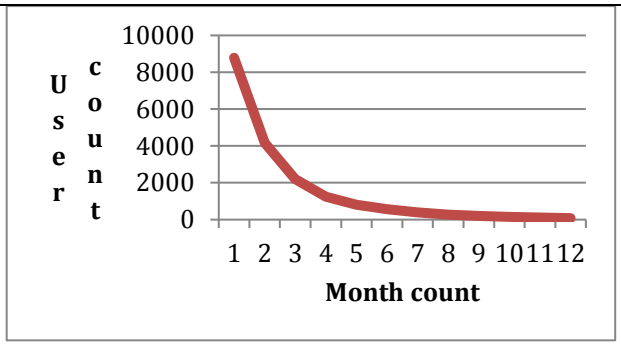

Figure 7.30: User count vs. Month count after removing users active for less than 3 month in "Pet supplies" category

B. Data clustering

Table below show detail statistics of all features for each cluster in terms of minimum, maximum, mean, and standard deviation values in all seven-product categories.

Electronics

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	7.502	4	261	6.080
C2	9.548	4	489	11.098
C3	9.197	4	178	8.1370
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	746.024	98	6978.75	551.94
C2	735.983	83.667	14690.333	632.023
C3	512.683	79.857	5374.833	397.094
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	2.968	1	3.75	0.561
C2	4.3908	3.375	5	0.406
C3	4.349	3	5	0.418
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	5.798	4	80	3.248
C2	6.268	4	105	4.478
C3	5.856	4	74	3.587
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	0.714	0	1	0.146
C2	0.852	0.704	1	0.080
C3	0.600	0	0.734	0.107

Table 7.8: Statistics of feature set in different clusters for "Electronics" category

Cellphones and accessories

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	6.784	4	54	3.527
C2	7.851	4	117	5.791
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	562.059	98.250	7351.333	494.669
C2	537.654	96.5	12010	560.721
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	3.211	1	4.5	0.556
C2	4.392	3.5	5	0.379
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	4.93	4	25	1.713
C2	5.143	4	38	2.22
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	0.678	0	1	0.169
C2	0.769	0	1	0.147

Table 7.9: Statistics of feature set in different clusters for "Cellphones and accessories" category

Health and personal care

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	6.234	3	295	7.948
C2	5.368	3	136	4.901
C3	6.684	3	297	7.362
C4	6.342	3	147	6.993
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	556.852	74	18288.337	475.565
C2	598.995	98.333	9027.75	455.338
C3	352.357	72	4954.5	261.466
C4	415.873	77.5	5731.5	311.301
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	4.569	3.75	5	0.366
C2	3.090	1	3.889	0.663
C3	4.664	4.05	5	0.295
C4	3.468	1	4.056	0.565
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	4.421	3	72	3.200
C2	4.221	3	58	2.603
C3	4.358	3	61	2.944
C4	4.477	3	52	3.421
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	0.632	0.43	1	0.143
C2	0.610	0.31	1	0.143
C3	0.238	0	0.443	0.129
C4	0.247	0	0.454	0.124

Table 7.10: Statistics of feature set in different clusters for "Health and personal care" category

Grocery and gourmet food

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	8.862	3	247	11.486
C2	7.359	3	513	13.893
C3	7.158	3	364	8.984
C4	7.245	3	224	9.813
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	428.211	43.4	4727.235	307.289
C2	473.124	82.667	6244.250	366.497
C3	318.656	74.333	6146.05	318.656
C4	491.464	75.333	6724.667	366.688
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	3.641	1	4.333	0.481
C2	4.596	3.333	5	0.381
C3	4.707	4.067	5	0.279
C4	3.24	1	4.176	0.666
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	5.944	3	57	5.282
C2	4.808	3	82	4.244
C3	4.583	3	60	3.324
C4	5.085	3	60	4.224
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	0.186	0	0.407	0.114
C2	0.654	0.417	1	0.148
C3	0.234	0	0.45	0.133
C4	0.515	0.167	1	0.132

Table 7.11: Statistics of feature set in different clusters for "Grocery & gourmet food" category

Office products

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	5.545	3	103	5.49
C2	4.266	3	60	2.409
C3	6.08	3	167	5.713
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	764.199	83	16312	699.689
C2	717.387	79.5	6517.333	576.397
C3	442.537	77.5	6326.533	360.158
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	4.398	3.25	5	0.473
C2	2.652	1	3.667	0.657
C3	4.368	2.25	5	0.553
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	4.448	3	51	3.268
C2	3.765	3	27	1.617
C3	4.657	3	57	3.500
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	0.661	0.438	1	0.158
C2	0.539	0	1	0.204
C3	0.213	0	0.445	0.136

Table 7.12: Statistics of feature set in different clusters for "Office product" category

Baby

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	8.193	4	40	4.13
C2	11.195	4	98	8.429
C3	9.602	4	89	6.126
C4	8.41	4	55	5.114
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	648.17	136.632	3109.95	385.613
C2	679.791	143.4	2827.429	357.196
C3	432.561	102.4	2684.375	236.848
C4	1035.95	192.077	6048.8	641.918
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	3.284	1.167	3.9	0.465
C2	4.394	3.625	5	0.322
C3	4.413	3.286	5	0.383
C4	3.579	2.154	5	0.524
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	5.272	4	22	1.779
C2	6.144	4	28	3.117
C3	4.939	4	15	1.413
C4	5.688	4	36	2.621
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	0.327	0	0.77	0.129
C2	0.589	0.32	0.73	0.83
C3	0.137	0	0.28	0.081
C4	0.68	0.45	1	0.117

Table 7.13: Statistics of feature set in different clusters for "Baby" category

Beauty

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	6.652	3	303	8.565
C2	6.497	3	368	8.485
C3	**7.149**	3	150	7.384
C4	6.954	3	241	8.005
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	523.868	78.33	7931	412.313
C2	**569.676**	89.667	8615	424.128
C3	332.453	67	4122.75	237.288
C4	395.392	77	3676.5	273.945
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	4.596	3.8	5	0.346
C2	3.274	1	4.009	0.601
C3	**4.654**	4.048	5	0.299
C4	3.452	1	4.071	0.555
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	4.326	3	82	2.884
C2	**4.366**	3	50	2.735
C3	4.137	3	44	2.277
C4	3.452	3	48	2.593
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	**0.644**	0.43	1	0.145
C2	0.605	0.36	1	0.14
C3	0.233	0	0.44	0.134
C4	0.239	0	0.44	0.128

Table 7.14: Statistics of feature set in different clusters for "Beauty" category

Pet supplies

Cluster	Mean total review count	Min total review count	Max total review count	Standard deviation
C1	6.187	3	192	6.237
C2	5.595	3	74	4.436
C3	6.459	3	136	5.385
C4	6.154	3	86	5.121
Cluster	**Mean average review length**	**Min average review length**	**Max average review length**	**Standard deviation**
C1	622.663	84.75	8277.5	464.249
C2	671.242	103.33	10188	519.079
C3	380.454	74.5	3208	257.954
C4	435.547	88	4049.333	302.973
Cluster	**Mean average overall**	**Min average overall**	**Max average overall**	**Standard deviation**
C1	4.519	3.75	5	0.373
C2	3.134	1	3.889	0.604
C3	4.663	4.063	5	0.293
C4	3.526	1	4.091	0.507
Cluster	**Mean total active month**	**Min total active month**	**Max total active month**	**Standard deviation**
C1	4.336	3	49	2.609
C2	4.118	3	33	2.144
C3	3.987	3	31	1.768
C4	4.003	3	35	1.971
Cluster	**Mean average helpfulness**	**Min average helpfulness**	**Max average helpfulness**	**Standard deviation**
C1	0.631	0.41	1	0.15
C2	0.613	0.3	1	0.154
C3	0.208	0	0.43	0.135
C4	0.214	0	0.43	0.13

Table 7.15: Statistics of feature set in different clusters for "Pet supplies" category

C. Data classification- J48 pruned decision tree

Figures below show top three levels of J48 decision trees for all seven-product categories.

Electronics

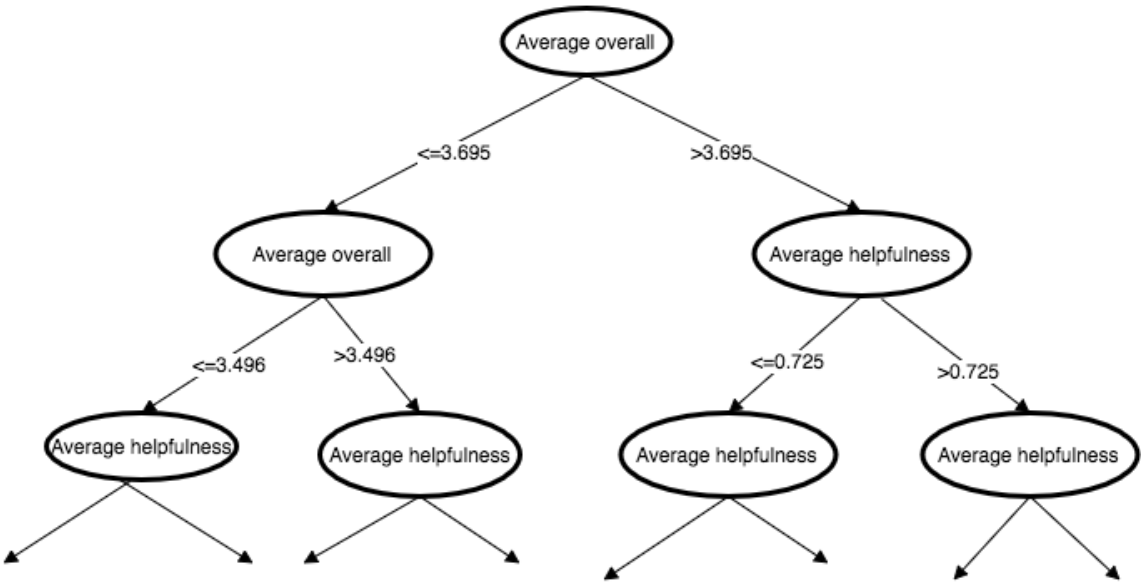

Figure 7.31: First three level of J48 pruned tree of "Electronics" category

Cellphones and accessories

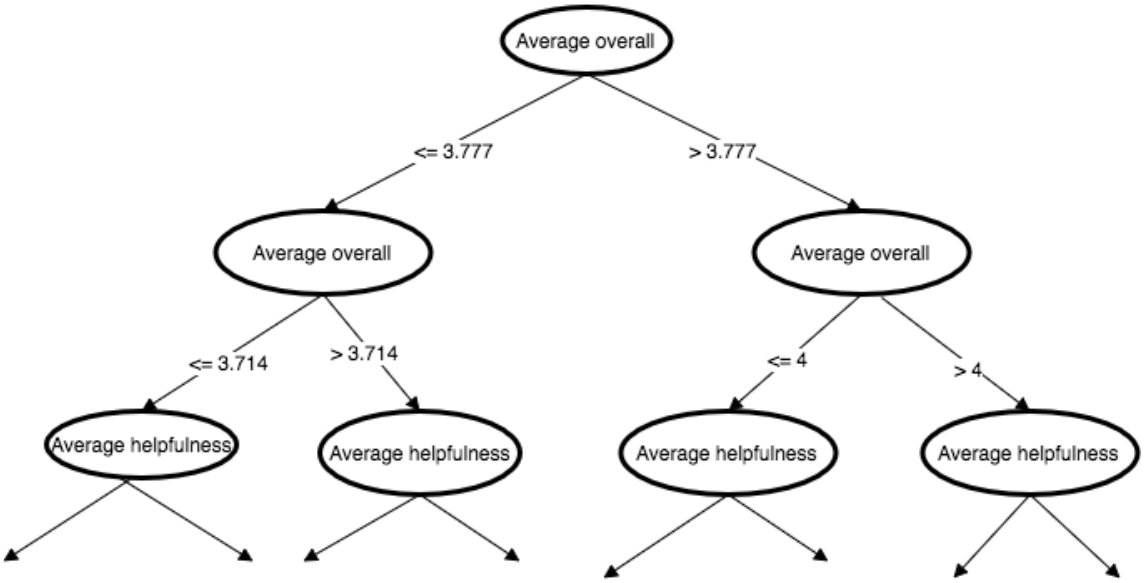

Figure 7.32: First three level of J48 pruned tree of "Cellphones and accessories" category

Health and personal care

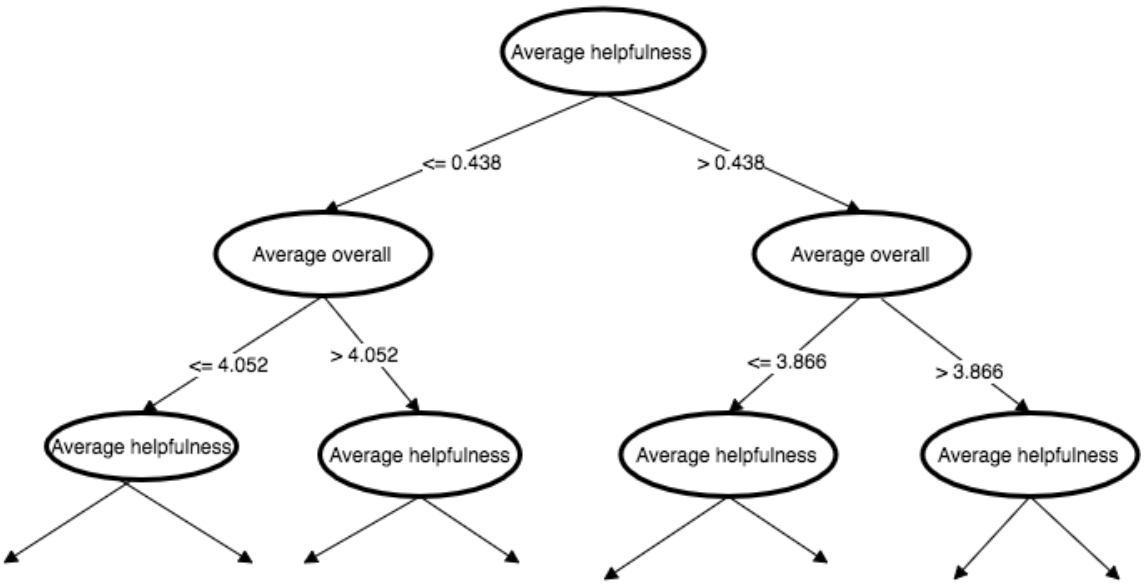

Figure 7.33: First three level of J48 pruned tree of "Health & personal care" category

Grocery and gourmet food

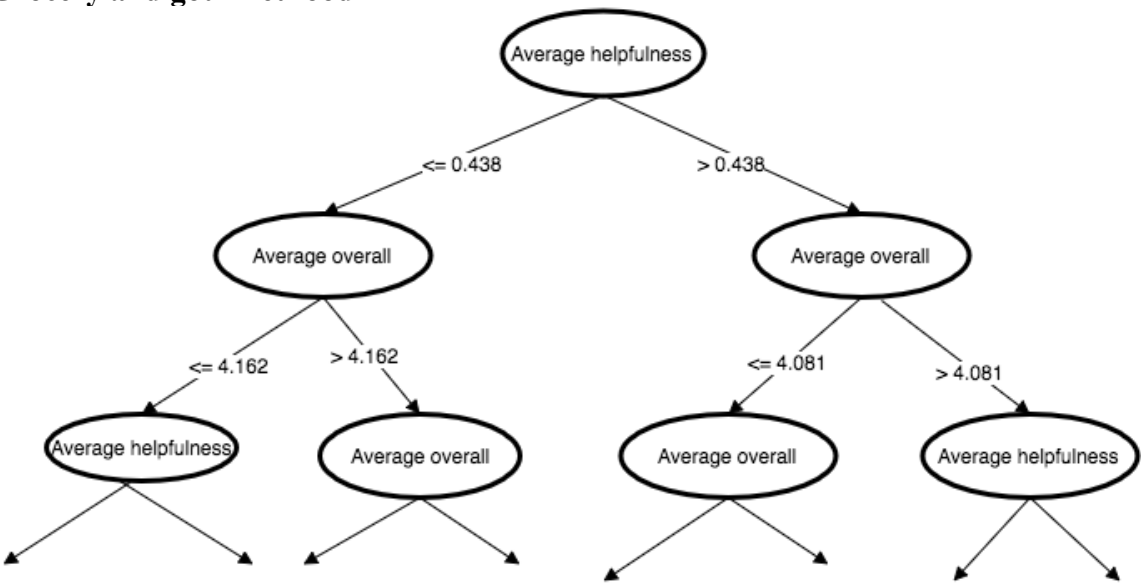

Figure 7.34: First three level of J48 pruned tree of "Grocery & gourmet foods" category

Office products

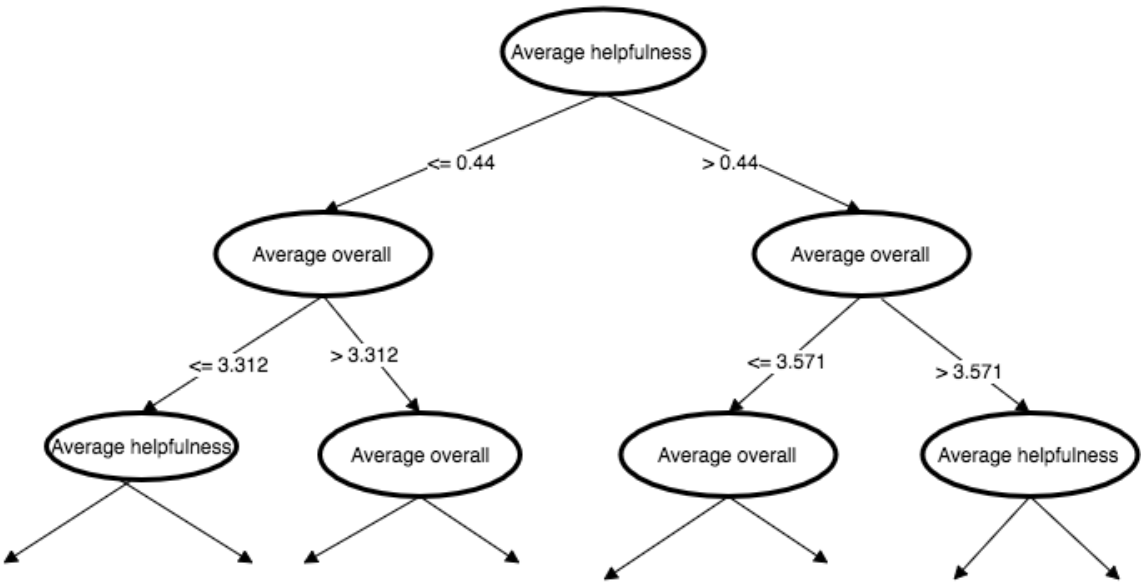

Figure 7.35: First three level of J48 pruned tree of "Office product" category

Baby

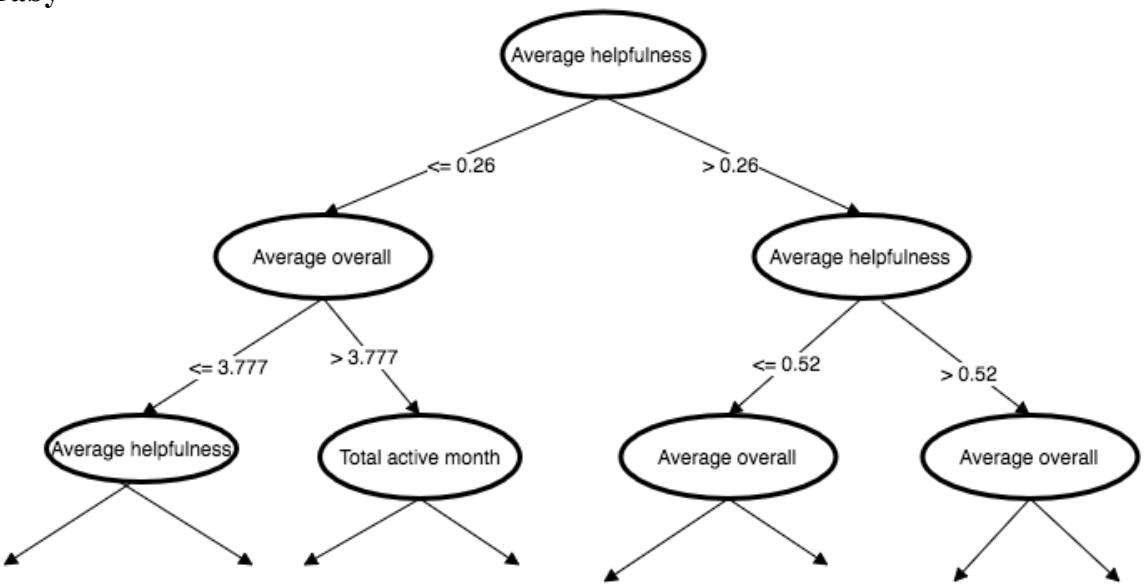

Figure 7.36: First three level of J48 pruned tree of "Baby" category

Beauty

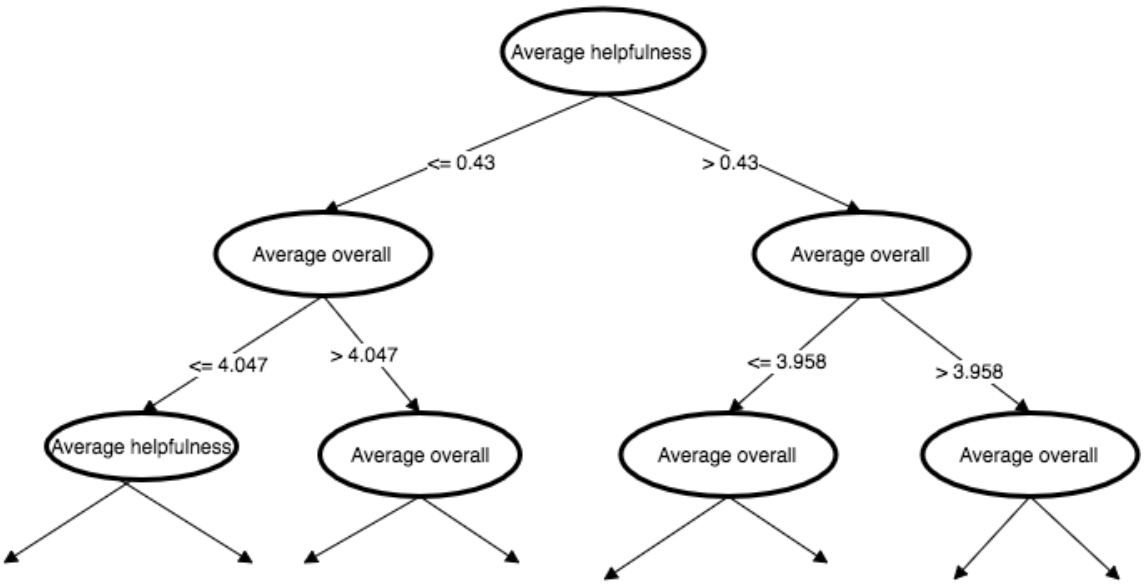

Figure 7.37: First three level of J48 pruned tree of "Beauty" category

Pet supplies

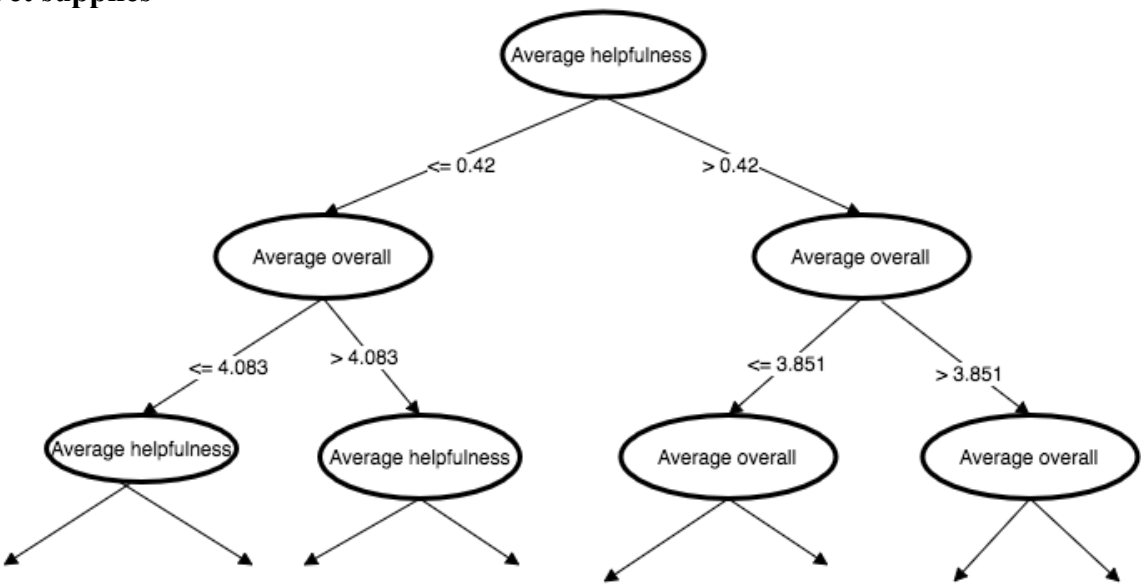

Figure 7.38: First three level of J48 pruned tree of "Pet supplies" category

D. Data classification- Confusion Matrix

Tables below show confusion matrices after performing10-fold cross validation in all seven-product categories.

Electronics

Classified as	Cluster1	Cluster2	Cluster3
Cluster1	33044	20	63
Cluster2	32	70375	88
Cluster3	59	82	42394

Table 7.16: Confusion matrix for "Electronics" category

Cell phones and accessories

Classified as	Cluster1	Cluster2
Cluster1	7706	34
Cluster2	36	13910

Table 7.17: Confusion matrix for "Cellphones and accessories" category

Grocery and gourmet food

Classified as	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	6699	0	30	27
Cluster2	0	9265	17	16
Cluster3	33	12	13472	4
Cluster4	22	17	7	5090

Table 7.18: Confusion matrix of "Grocery & gourmet food" category

Health and personal care

Classified as	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	20251	18	12	3
Cluster2	7	9697	0	15
Cluster3	15	0	25090	2
Cluster4	4	19	1	15535

Table 7.19: Confusion matrix of "Health & personal care" category

Office product

Classified as	Cluster1	Cluster2	Cluster3
Cluster1	12998	3	14
Cluster2	7	6696	13
Cluster3	14	14	3453

Table 7.20: Confusion matrix of "Office product" category

Baby

Classified as	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	1104	14	11	12
Cluster2	15	1792	30	36
Cluster3	7	25	1967	0
Cluster4	13	41	0	966

Table 7.21: Confusion matrix of "Baby" category

Beauty

Classified as	Cluster1	Cluster2	Cluster3	Cluster4

Cluster1	12116	16	5	1
Cluster2	14	6702	0	36
Cluster3	9	0	6702	8
Cluster4	0	13	6	9097

Table 7.22: Confusion matrix of "Beauty" category

Pet supplies

Classified as	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	7707	6	10	7
Cluster2	7	4021	0	12
Cluster3	5	0	11380	8
Cluster4	7	13	4	6529

Table 7.23: Confusion matrix of "Pet supplies" category

E. User Evolution

Here we report on our change analysis of helpfulness for all seven-product categories. For each class, we calculate the average helpfulness of all reviews written by reviewers belonging to the particular class with respect to time (in year). The average helpfulness of all classes of reviewers in all seven-product categories are drawn below.

Electronics

There are three classes of reviewers in Electronics: 1) conscientious, 2) expert, and 3) novice.

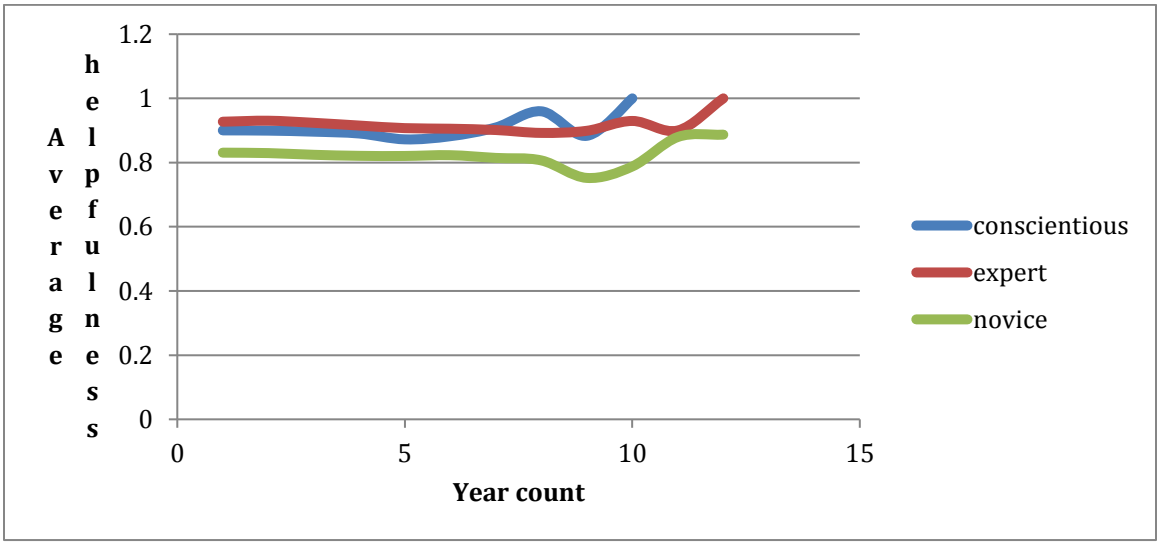

Figure 7.39: Trend of average helpfulness over time for three clusters in "Electronics" category

Cellphones and accessories

There are two classes in Cellphones and accessories: 1) conscientious, and 2) expert. For

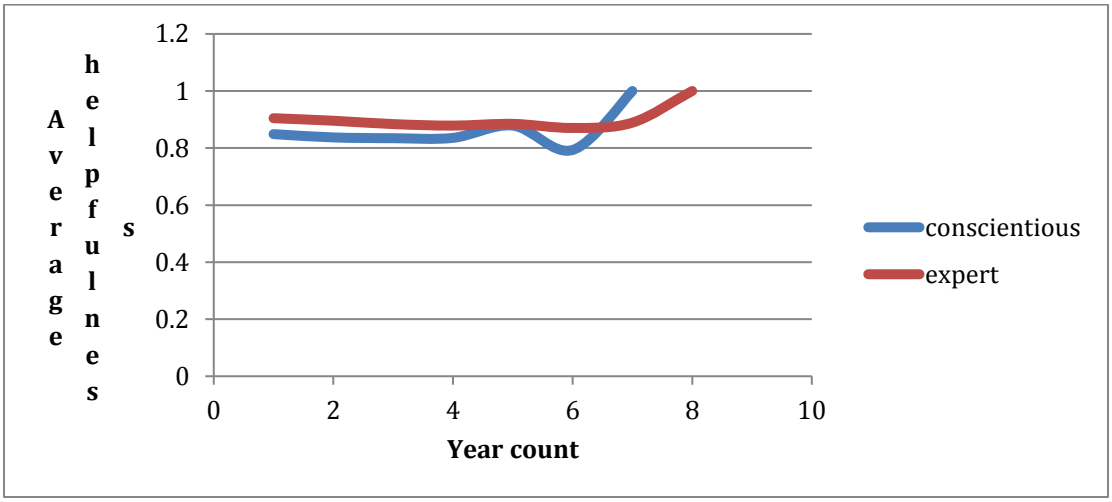

Figure 7.40: Trend of average helpfulness over time for three clusters in "Cellphones and accessories" category

Office product

There are three classes of reviewers in Office products: 1) conscientious, 2) expert, and 3) novice.

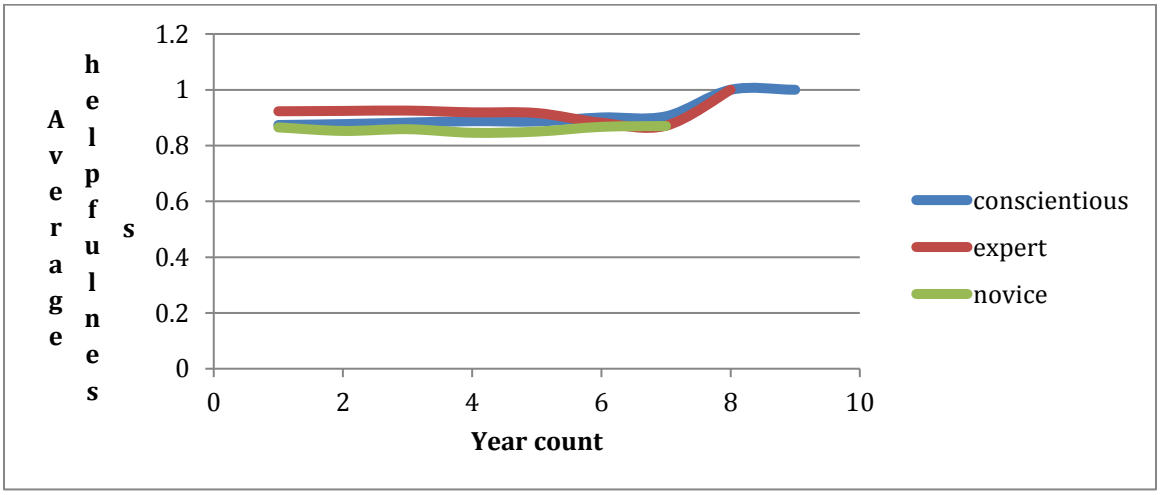

Figure 7.41: Trend of average helpfulness over time for three clusters in "Office product" category

Grocery and gourmet food

There are four classes of reviewers in Grocery and gourmet foods: 1) novice, 2) conscientious, 3) positive expert, and 4) negative experts.

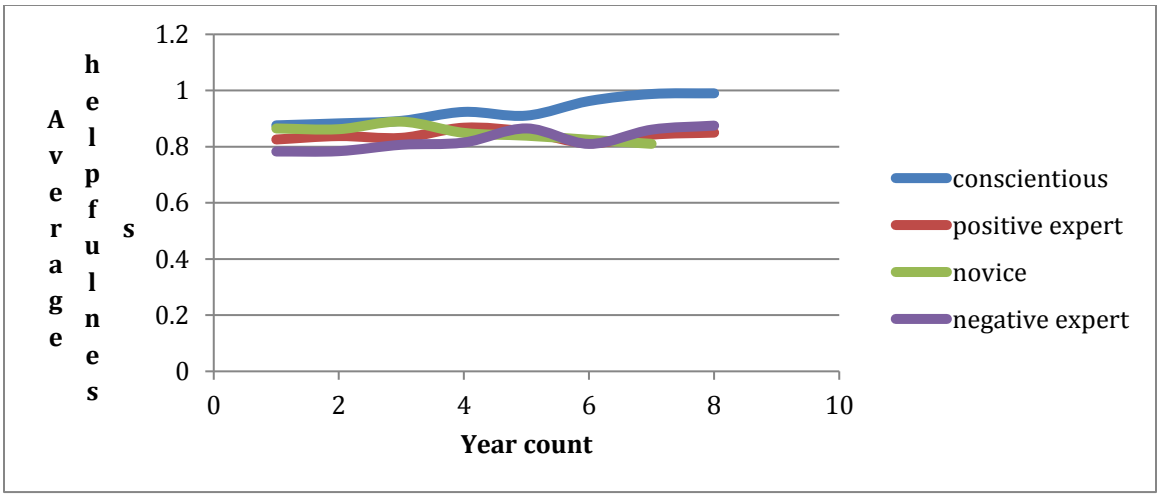

Figure 7.42: Trend of average helpfulness over time for four clusters in "Grocery and gourmet food" category

Health and personal care

There are four classes of reviewers in Health and personal care: 1) novice, 2) conscientious, 3) frequent positive expert, and 4) non-frequent negative experts.

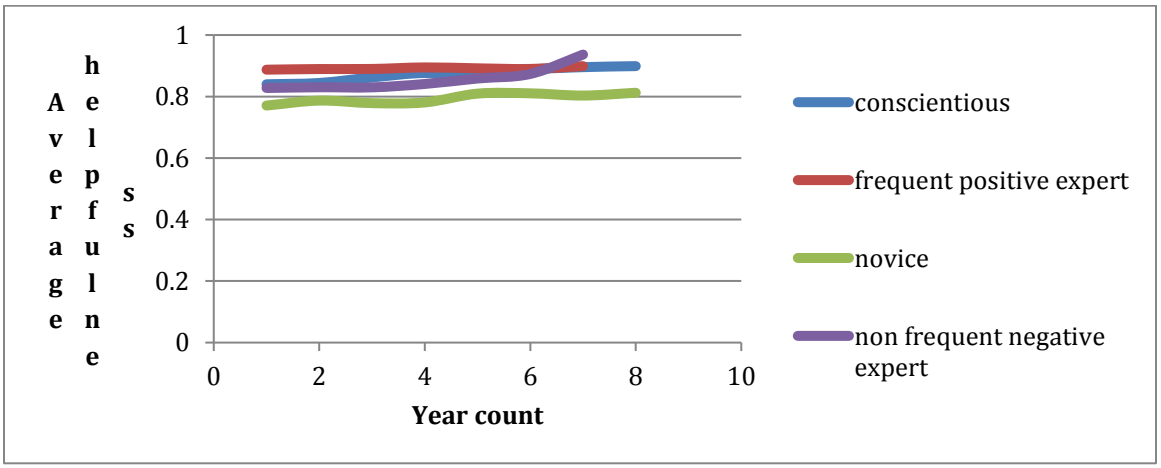

Figure 7.43: Trend of average helpfulness over time for four clusters in "Health and personal care" category

Baby

There are four classes of reviewers in Baby: 1) novice, 2) conscientious, 3) frequent positive expert, and 4) non-frequent negative experts.

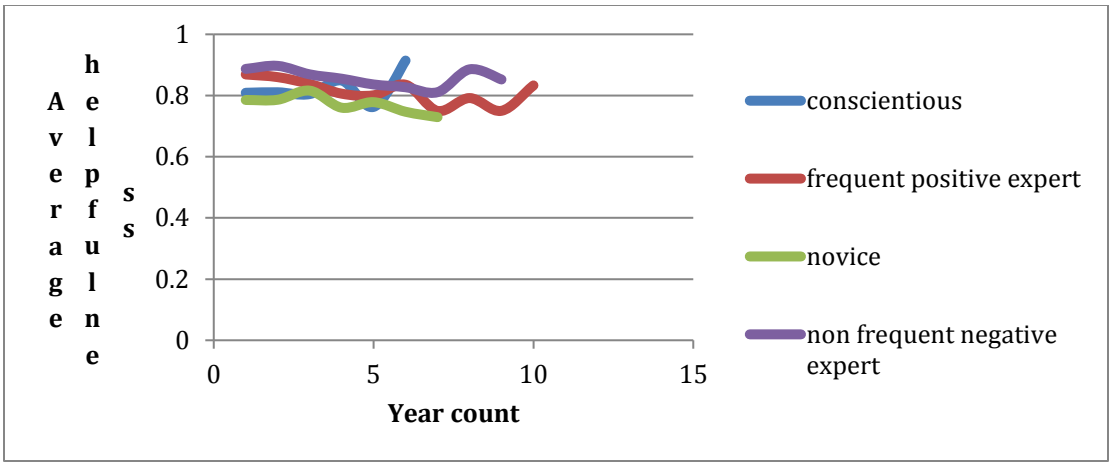

Figure 7.44: Trend of average helpfulness over time for four clusters in "Baby" category

Beauty

There are four classes of reviewers in Beauty: 1) novice, 2) conscientious, 3) positive expert, and 4) negative expert.

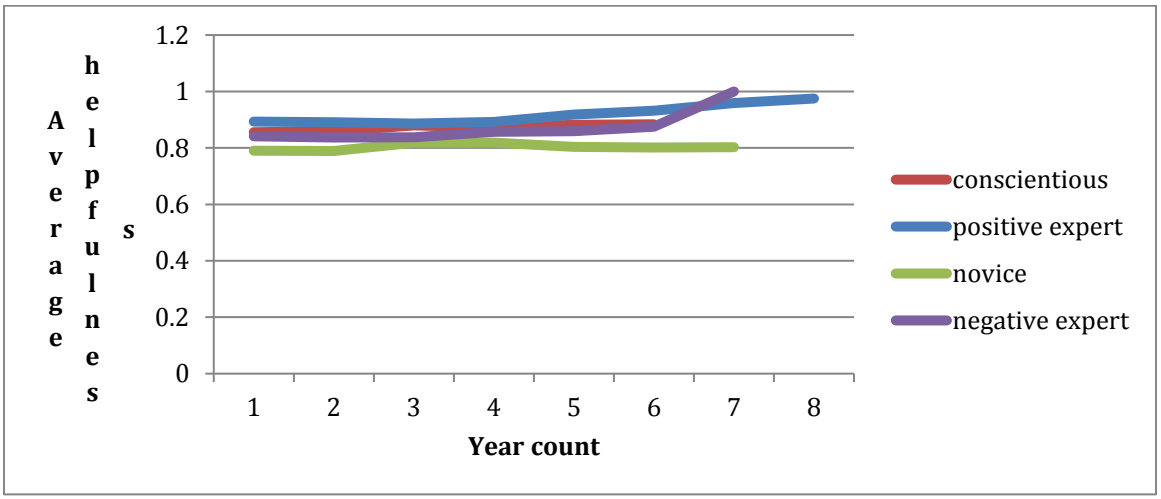

Figure 7.45: Trend of average helpfulness over time for four clusters in "Beauty" category

Pet supplies

There are four classes of reviewers in Pet supplies: 1) novice, 2) conscientious, 3) frequent positive expert, and 4) non-frequent negative experts.

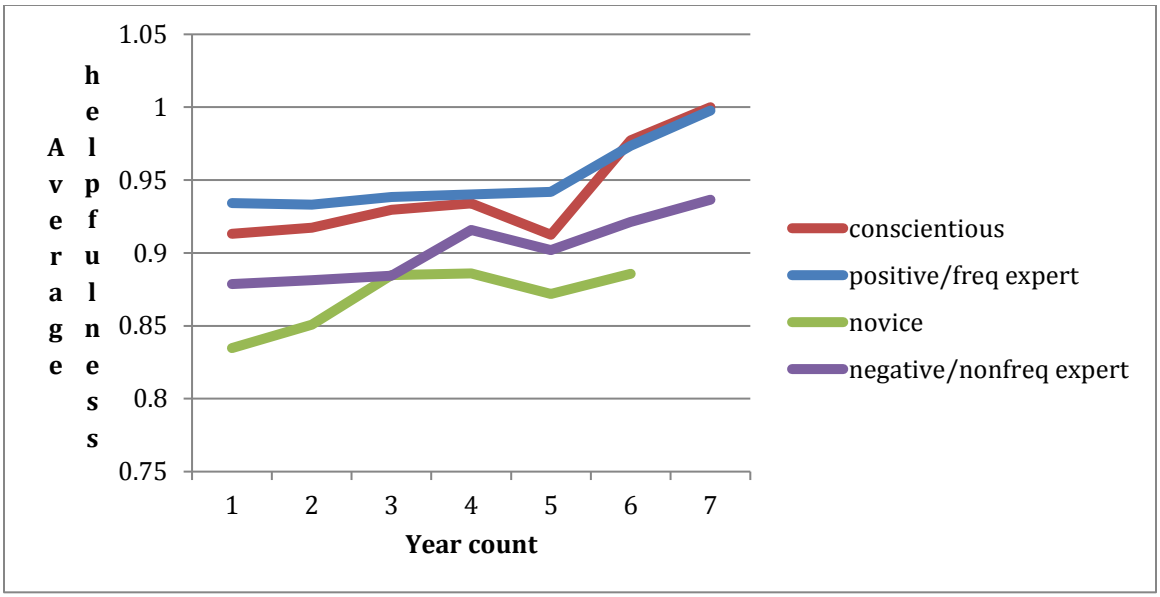

Figure 7.46: Trend of average helpfulness over time for four clusters in "Pet supplies" category

F. Sentiment Analysis

In this Section, for each reviewer class, we calculate correlation between review features such as average helpfulness, and average overall with review sentiments measured by positive intensity, negative intensity, and neutral intensity.

		Helpfulness			Overall		
Product Category	**Clusters**	**Positive**	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**
Books	C1 (conscientious)	-0.292	0.153	0.200	0.364	-0.373	-0.138
	C2 (expert)	-0.288	0.111	0.243	0.243	-0.212	-0.145
	C3 (novice)	-0.239	0.134	0.187	0.201	-0.197	-0.117
Cellphones & accessories	C1 (conscientious)	-0.078	0.000	0.084	0.407	-0.421	-0.258
	C2 (expert)	-0.061	-0.037	0.082	0.415	-0.450	-0.220
Electronics	C1 (novice)	-0.085	0.017	0.067	0.338	-0.3169	-0.1140
	C2 (expert)	-0.112	0.056	0.088	0.310	-0.309	-0.172
	C3 (conscientious)	-0.091	0.004	0.091	0.277	-0.273	-0.164
Office product	C1 (expert)	-0.114	0.054	0.091	0.267	-0.273	-0.149
	C2 (novice)	-0.013	-0.012	0.021	0.369	-0.319	-0.120
	C3 (conscientious)	-0.077	0.019	0.071	0.311	-0.292	-0.191
Grocery & gourmet food	C1 (conscientious)	-0.083	0.047	0.068	0.195	-0.194	-0.129
	C2 (novice)	-0.081	0.015	0.072	0.314	-0.313	-0.148
	C3 (positive expert)	-0.076	0.011	0.073	0.224	-0.193	-0.154

	C4 (negative expert)	-0.051	0.005	0.046	0.348	-0.322	-0.154
Health & personal care	C1 (frequent positive expert)	-0.049	0.029	0.034	0.200	-0.129	-0.136
	C2 (non frequent negative expert)	-0.056	-0.010	0.0595	0.300	-0.252	-0.116
	C3 (conscientious)	-0.083	0.060	0.056	0.174	-0.117	-0.123
	C4 (novice)	-0.043	0.036	0.019	0.258	-0.266	-0.088
Baby	C1 (novice)	-0.128	-0.001	0.128	0.338	-0.267	-0.195
	C2 (frequent positive expert)	-0.039	0.029	0.029	0.263	-0.293	-0.155
	C3 (conscientious)	-0.069	-0.004	0.075	0.361	-0.351	-0.246
	C4 (non frequent negative expert)	-0.016	0.0009	0.017	0.363	-0.373	-0.198
Beauty	C1 (positive expert)	-0.2158	0.6379	-0.7200	0.3111	-0.4441	0.4241
	C2 (negative expert)	-0.1665	0.8632	-0.7726	0.3404	-0.2955	0.1984
	C3 (conscientious)	-0.2056	0.8623	-0.6150	0.2976	-0.6270	0.3670
	C4 (novice)	-0.1726	0.8180	-0.7357	0.3465	-0.3670	0.2517
Pet supplies	C1 (frequent positive expert)	-0.1872	0.8630	-0.7534	0.3073	-0.5522	0.3910
	C2 (non frequent negative expert)	-0.1466	0.8490	-0.8377	0.3477	-0.2980	0.2223
	C3 (conscientious)	-0.1863	0.8700	-0.6836	0.2722	-0.6377	0.4302
	C4 (novice)	-0.1516	0.8042	-0.77892	0.3053	-0.3899	0.3093

Table 7.24: Correlation between helpfulness and overall with respect to positive, negative, and neutral tone for each cluster.

G. Class definition

This Section contains the definition of review and reviewer class.

Review class

```
Class Review
{       String author;                  // person who posted the review
        String helpfulness;             //current helpfulness of the review
```

```
        Integer score;      // add 1 for successful and -1 for unsuccessful recommendation
        Integer recommendationCount;                // number of times the review has been
                                                    // recommended
        Double rank;                                        // denotes success rate of review
}
```

Reviewer class

```
Class Reviewer
{       Double avgHelpfulness;              // average helpfulness of the reviewer
        Double previousAvgHelpfulness;      // average helpfulness of the reviewer before
                                            // reviewer was recommended

        List avgHelpfulnessList;            // average helpfulness of the reviewer for
                                            // each active year
        Integer score;      // add 1 for successful and -1 for unsuccessful recommendation
        Integer recommendationCount; // number of times a review posted by reviewer
                                                    // has been recommended
        Double rank;                                // denotes success rate of reviewer
}
```