MINING AND SUMMARIZING CUSTOMER REVIEWS BY GENERATING FEATURE-SPECIFIC RATINGS

Prabin Lamichhane

MINING AND SUMMARIZING CUSTOMER REVIEWS BY GENERATING FEATURE-SPECIFIC RATINGS

by

PRABIN LAMICHHANE

Presented to the Faculty of the Honors College of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

HONORS BACHELOR OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2023

ACKNOWLEDGMENTS

I like to extend my sincere gratitude to my faculty mentor Dr. Chengkai Li for providing guidance and suggestions throughout the project implementation. Also, thanks go to Dr. Vassilis Athitsos for taking the time to revise my work. His teaching materials were the starting point for my exploring Bayesian techniques. Along the way, I would like to acknowledge Dr. Victoria Chen for providing feedback on my work. Finally, a massive acknowledgment goes to those who contributed to open-source natural language processing libraries in Python. Without them, this project would need a lot more work.

November 17, 2022

ABSTRACT

MINING AND SUMMARIZING CUSTOMER REVIEWS

BY GENERATING FEATURE-SPECIFIC

RATINGS

Prabin Lamichhane, B.S. Computer Science

The University of Texas at Arlington, 2023

Faculty Mentor: Chengkai Li

This paper proposes a novel method to generate ratings from reviews using a Bayesian technique. One of the reasons for the growing trend of online shopping in e-commerce platforms is its transparent review system, where a customer can review and rate a product that becomes open for others to see. Oftentimes, in making a purchase decision, a customer reads these reviews to get feature-specific information about a product. These reviews, however, are becoming increasingly incomprehensible for a person to read in their entirety because of their large volume. Reading a sample of them may create a biased opinion as they do not represent overall reviews. To solve this problem, this project used Bayesian estimation to develop fine-grained, feature-specific ratings of products from the reviews of customers. This task is performed in three steps: (1) mining product features from the reviews of customers (2) identifying the sentiment of the reviews that describe product

features (3) generating feature-specific ratings in 5-Point Likert scale. The ratings are generated using the Bayesian approach and are compared with the ones generated using the Frequentist approach.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation and challenges

The motivation for this research stems from a need to make the growing amount of information in e-commerce easily accessible. In e-commerce platforms like Amazon, customers can publicly share their experience of the products they bought and give a rating, usually on a 5-Point Likert Scale (see Appendix A). These responses are helpful for other customers in making a purchase decision. Also, they become a critical source of information for sellers and manufacturers to improve their products.

However, with the rise in customers shopping online, reviews are growing rapidly. For most of the products online today, these reviews are in the order of thousands. It makes it harder for customers to read and comprehend such a high volume of information. On top of that, businesses may highlight the top reviews and sway the customers to buy the product.

In addition, the product ratings itself does not give much information about a product. They are subject to biases from personal interests. For instance, customer A may prioritize the color of a product and rate 5-point, whereas customer B may like the sound of the product and rate 5-point. Here, the rating 5 does not give any interpretive information because they are subject to personal interests of reviewers. Customers may, however, want to know about the features of a product and make an informed decision based on what they are interested in. This project addresses these challenges in comprehending large volume of reviews by using natural language processing techniques to extract feature-specific information about a product from customer reviews and generate ratings. However, reviews that talk about a specific feature may be relatively small.

There are generally two approaches to generate ratings. Using a Frequentist approach to generate ratings in a small dataset may create a biased result because it strictly considers only the ratings it has seen and not the unobserved possibilities. [1] A better-suited approach is the Bayesian estimation, where prior-belief and likelihood distribution estimate the posterior belief.[2] Prior belief represents the information about the product ratings before seeing any reviews, and likelihood distribution tells the probability of seeing a positive or negative review if the rating is known. Posterior belief highlights how evidence updates the prior belief based on the likelihood of the evidence. Such a method would produce a robust rating that is less influenced by a small set of reviews.

1.2 Literature reviews

Some of the related past works summarize the number of positive and negative reviews of the product features that appear frequently in the reviews. Hu et al.[3] use a similar approach to summarize the product features. Their approach can be broken into three steps: 1) mining product features from the reviews 2) identifying opinion sentences in the reviews and classifying them as positive or negative 3) summarizing the opinions as the number of positive reviews and negative reviews received about that feature. This project's approach is similar to the one proposed by Hu et al. but differs in a few aspects. Their approach of summarizing the reviews based on the number of positive or negative reviews leads to biased interpretation as different people might interpret these numbers differently. This project's approach generates 5-Point Likert Scale ratings from these numbers, which is intuitive to understand for anybody.

Other related works follow a similar approach to Hu et al. but generate text summarization of the reviews as an output. Wang et al.[4] first mine the reviews, extract opinion sentences from the reviews, and generate review summaries using a feature-based weighted non-negative matrix factorization method (FNMF). This project's approach is different in the sense that our summary of the reviews is based on ratings rather than on textual summaries. One problem with using a text summarization of the product features is that it can be biased as the summary of the product feature could be perceived differently by different people.

CHAPTER 2

METHODOLOGY

2.1 Framework

The overall methodology implements a conceptual framework that consists of four sequential steps as in the figure below.

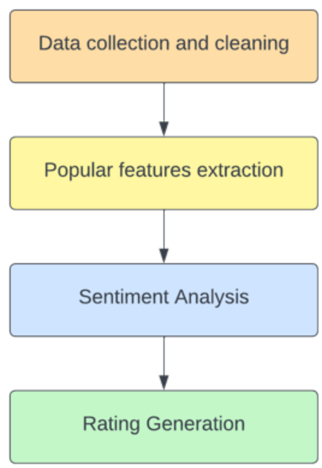

Figure 2.1: Conceptual framework of rating generation from customer reviews

2.2 Data collection and cleaning

A freely available Amazon product review dataset was collected from Kaggle.[5] There were a total of 1599 reviews for 62 different products. The features, such as reviews and product, were extracted. Any data point that had empty or null value in reviews and product column were filtered out as part of the data cleaning.

2.3 Popular features extraction

The feature extraction process started with identifying potential features. Usually, these features are nouns. So, the goal was to extract nouns from the reviews. One caveat with this approach is that it only considers those features that have been explicitly mentioned but do not capture the ones that have been implied. For example, consider a review sentence "The device was small" about the size of a product. From this sentence, it can be implied that the device size was small. However, it would not be considered during feature extraction because it does not explicitly mention the word 'size' or any specific nouns that describe the small feature. Nevertheless, this approach of feature extraction captures majority of the features in the review sentences. Next, the features that frequently appeared in the reviews were identified. These frequent features are the popular features.

To do the feature extraction, customer reviews were broken into sentences using a sentence tokenizer of the **NTLK** library in Python. These sentences were stored in a separate sentence database. Then, each of these sentences was passed through a Part-Of-Speech (POS) tokenizer that parsed each word in the sentence and identified what part of speech it belonged to. Then, all the nouns and noun phrases were extracted from these sentences. These extracted nouns and noun phrases are the potential features.

Next, the goal was to identify popular features. A transaction file was created, where each transaction contained nouns from a review sentence. Then, the Apriori algorithm[6] was implemented with the support of 1% to generate the popular features. 1% support means that if a noun appears in at least 1% of the transactions, it is a popular feature. Then, from the list of the top 100 popular nouns and noun phrases, the non-features were filtered out manually. The filtered result contained the list of popular features.

2.4 Sentiment analysis

Then, each of the review sentences from the above step was classified into positive, negative, or neutral using a rule-based sentiment analyzer [7]. This project's approach is to estimate ratings based on how positive or negative a review sentence is. The neutral sentences do not contain strong opinions, which is no dominant inclination towards either positive or negative sentiment, about the feature, so the neutral sentences do not contribute towards the rating. Therefore, for computational efficiency, they were removed from the sentence database created above. At this stage, the sentence database contained the product names, review sentences, a list of popular features in that review sentence, and the sentiment of the review sentence as positive or negative.

2.5 Rating generation

The next task was to generate ratings for each popular feature of a product. For each product, all the review sentences that contained at least one of the popular features as nouns were extracted. For each of those features, a list of sentiments of the review sentence was created. If a review sentence contained a feature, its sentiment was added to the list of that feature. These observations of positive and negative sentences were used to generate ratings.

$Let\ S = \{s_1, s_2, s_3, \ldots, s_n\},$

$be\ the\ observations\ of\ review\ sentences\ that\ talks\ about\ a\ specific\ feature\ of\ a\ product,$

$where\ s_i is\ either + ve\ or - ve$

$Let\ J = \{1, 2, 3, 4, 5\}$

$Let\ R = \{r_i\}^J,\ be\ the\ ratings\ for\ that\ feature\ of\ a\ product, where\ r_i = J_i$

2.5.1 Frequentist approach

Frequentist way of computing the rating is to average over all the observations. A positive sentence observation was considered to be of rating 5 and that of a negative sentence observation was rating 1. Then, the expected value was computed with each of the observation being equally likely. This expected value is the rating of that feature of a product. Mathematically,

$$No\ of\ +ve\ review\ sentences = \left||S|\right|, where\ s_i\ is\ positive$$

$$No\ of\ -ve\ review\ sentences = \left||S|\right|, where\ s_i is\ negative$$

$$Total\ sentences = No\ of\ +ve\ review\ sentences + No\ of\ -ve\ review\ sentences$$

$$E[R] = \frac{No\ of\ +ve\ review\ sentences}{total\ sentences} * 5 + \frac{No\ of\ -ve\ review\ sentences}{total\ sentences} * 1$$

2.5.2 Bayesian approach

According to the Bayesian approach, prior belief about the ratings of a feature of a product is stated. This belief is represented as a uniform distribution, where each of the ratings are equally likely. This represents a customer's state of mind when he/she has no other information about the rating of a feature of a product. Then, a likelihood distribution was created that tells how likely it is to observe a positive or negative review sentence given the rating. Using these priors and likelihood, a posterior distribution of rating is generated based on the observation. Then, the posterior of the previous observation is used as the prior for the next observation. This way, an iterative algorithm will generate the posterior distribution having observed all the observations. Here, each of the positive or negative sentence observation is assumed to be conditionally independent of each other given the ratings. This assumption makes an intuitive sense because the likelihood of an observation does not change given the rating is known. The expected value of the posterior probability distribution is the generated rating. Mathematically,

Prior:

$P(r_i) = \{0.2, 0.2, 0.2, 0.2, 0.2\},$

$where\ the\ value\ at\ i^{th} index\ represents\ probability\ of\ rating\ r_i\ prior\ to\ any\ observation$

Likelihood:

$P(s_k =\ +ve\ |r_i) = \{0.1, 0.3, 0.5, 0.7, 0.9\},$

$where\ the\ value\ at\ i^{th} index\ represents\ the\ probability\ of\ observing\ a\ sentence$

$+\ ve\ given\ a\ rating\ r_i$

$P(s_k =\ -ve|r_i) = 1 - P(s_k =\ +ve|r_i)$

Posterior:

$P(r_i|s_1, s_2, \ldots, s_n) = \frac{P(s_n|r_i\ , s_1, s_2\ , \ldots, s_{n-1}) * P(r_i|s_1, s_2, \ldots, s_{n-1})}{\sum_r^R P(s_n|r, s_1, s_2, \ldots, s_{n-1}) * P(r|s_1, \ldots, s_{n-1})},$

$where\ P(s_n|\ r_i, s_1, s_2, \ldots, s_{n-1}) =\ P(s_n|r_i),$

$because\ of\ conditional\ independence\ assumption\ of\ s_i's\ given\ r_i$

Expected Rating:

$E[R|S] =\ \sum_r^R P(r|s_1, s_2, \ldots, s_n) * r$

CHAPTER 3

RESULTS

Features	Number of positive sentences	Number of negative sentences	Frequentist rating	Bayesian rating
Battery	3	0	5.0	4.45
Content	55	11	4.33	4.85
Quality	40	0	5.0	4.99

Table 3.1: Feature-specific ratings summary of Amazon Fire TV

The following results show that for a battery feature of Amazon Fire TV, there were a total of three positive sentences and zero negative sentences. In having observed just three positive sentences, the ratings computed using the Frequentist approach gives a 5-star rating. However, ratings computed using the Bayesian approach is more considerate of unknown possibilities and rates 4.45. It accounts for the prior belief and the likelihood of seeing the positive observation. This shows that ratings computed using Bayesian approach is not substantially dominated by a small number of reviews.

On the other hand, the rating of Quality feature is 4.99 according to the Bayesian. This rating is almost the same as the Frequentists rating of 5.0. From the observation, it is intuitively obvious that the rating generated by observing 40 positive sentences and zero negative sentences should be close to 5.0. This is observed in ratings generated using both approaches. This shows that for a large number of observations, the ratings computed using the Bayesian approach converge to the ones generated using the Frequentist approach.

Finally, the rating of Content is 4.85 according to Bayesian, and 4.33 according to the Frequentist. The decision about which rating makes more sense is a subjective question and should be made by gathering customers opinions through randomized study. However, having observed 55 positive ratings and 11 negative ratings, both ratings agree that the rating is higher, which is intuitive.

CHAPTER 4

CONCLUSION AND FUTURE WORK

The results support the conclusion that the ratings generated using the Bayesian approach are more robust to a small number of reviews and converge to the ratings generated using the Frequentist approach for a large number of reviews for a chosen prior and likelihood. Therefore, the Bayesian approach is more suited to generate ratings from the reviews than the Frequentist approach.

In the future, the following areas could be explored:

1. New methods of extracting implied features from the reviews could be used that do not contain features as explicit nouns.
2. Parametric Bayesian methodologies could be tested for ratings generation, and its effectiveness could be verified through randomized study.
3. Different priors and likelihood distribution could be experimented to generate ratings.

APPENDIX A

5-POINT LIKERT SCALE

A Likert Scale is an evenly spaced scale that shows how much a person agrees or disagrees with a particular statement or question. In a 5-Point Likert scale, the points range from 1 to 5, where 5 means high agreement and 1 means high disagreement. The ratings are usually collected in this scale. In an e-commerce setting, it can be used to analyze whether the customer is pleased or have issues with a product or service.

REFERENCES

1. Gelman, A. (2022). Not frequentist enough. https://statmodeling.stat.columbia.edu/2022/10/05/not-frequentist-enough
2. Shah, A. (2020). Understand Bayes Rule, Likelihood, Prior and Posterior. https://towardsdatascience.com/understand-bayes-rule-likelihood-prior-and-posterior-34eae0f378c5
3. Minqing, H. Bing, L. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA, 168–177.
4. Dingding, W. Shenghuo, Z. Tao, L. (2013). SumView: A Web-based engine for summarizing product reviews and customer opinions, Expert Systems with Applications.
5. Yasser, M. Amazon Product Reviews Dataset. https://www.kaggle.com/datasets/yasserh/amazon-product-reviews-dataset?resource=download
6. Agrawal, R. Srikant, R. (1994). Fast Algorithms for Mining Association Rules. IBM Almaden Research Center, San Jose, CA, USA.
7. Pocs, M. (2020). Lovecraft with Natural Language Processing – Part 1: Rule-Based Sentiment Analysis. https://rb.gy/avu20o

BIOGRAPHICAL INFORMATION

Prabin Lamichhane is an undergraduate student majoring in Computer Science at the University of Texas at Arlington (UTA). His intended graduation is in Spring 2023 with an Honors Bachelor of Science in Computer Science. His academic career consisted of involvement in the Honors College, participation in competitions and hackathons, involvement in research, a conference presentation, a leadership role at the Academic Success Center, and two software engineering internships as a full-stack developer intern and automation engineer intern. He has won the Dean's Engineering Challenge 2020, and secured 3rd place in Innovations Day 2022 poster presentation at UTA. His research experience includes developing simulation models for police patrolling operations of Arlington Police Department (APD) and for primary health care operations at John Peter Smith (JPS) hospital in Fort Worth, Texas. He has also presented his former research at the INFORMS conference 2022. His interest is in Machine Learning, Statistics, and Distributed systems. His future plans are to gain industry experience as a software engineer, pursue PhD to work on some real-world problems in Machine Learning, and work as a researcher.