Proposal

Data Mining and Predictive Modeling of Amazon Customer Reviews

Technical Points of Contact:

Name: W. Grant Hatcher
Email Address: whatch2@students.towson.edu
Mailing Address: Department of Computer and Information Sciences
Towson University, 7800 York Road, Towson, MD 21252

Name: Kevin McNamara
Email Address: kmcnamara@towson.edu
Mailing Address: Department of Marketing
Towson University, Stephens Hall, Towson, MD 21252

Period of Performance: 10/4/2018 - 12/6/2018 (2 months)

Executive Summary

With 43.5% market share in 2017, Amazon is the leader in online retail in the US. The tech giant possesses a wealth of data regarding customer purchases, preferences, reviews, and more. Using data mining techniques, we can harness this data to research patterns in customer sentiments and buying habits. In this proposal, we plan on exploring a set of Amazon customer reviews to analyze patterns between various attributes and develop a predictive tool for discovered correlated variables. Considering a dataset of over 130 million reviews provided from Amazon's AWS service, we have the potential to consider the impacts of positive and negative reviews on product ratings and customer sentiment. Moreover, applying advanced machine learning systems, such as deep neural networks in TensorFlow, we have the potential to provide high-accuracy predictive capabilities to be leveraged for consumer targeting, increased customer satisfaction, and the generation of additional revenue.

Contents

1. Introduction

1.1. Problem Statement

As a massive company subsuming large portions of a variety of consumer markets, Amazon generates a considerable amount of Big Data. This data represents a new type of natural resource in the modern digital era. What's more, the question of how to analyze and extract business insights from this massive data remains unresolved. The current dataset contains over 130,000,000 individual Amazon product reviews covering a range of 46 product categories, and containing attributes such as Customer ID, Product ID, Star Rating, Helpful Votes, Total Votes, and Review Date. The dataset, while detailed, does not alone relate reviews to one another, correlate attributes to show patterns, or provide clear intelligence on the particular types of reviews or reviewers that potential customers find helpful. These analytical measures must be achieved through the process of Data Mining must consider.

1.2. Challenges

A variety of challenges exist in considering such Big Data, defined by the Volume, Value, Velocity, Veracity, etc. that make such data too large for traditional data processing techniques to handle. As a critical challenge, the size of the data will particularly make cleaning and processing time-consuming and difficult. In addition, despite the volume of data, it is unclear what, if any, value can be extracted. For instance, the purchasing history of the products would be an ideal additional data feature to have, but this is not present. Instead, we can only guess that certain types of reviews and certain trusted reviewers play a role in the positive purchasing experience of other customers.

1.3. Objectives

In our analysis of the Amazon Customer Review dataset, we consider several key objectives. Namely, we are interested in:

1. Conducting and exploratory analysis of the dataset to establish a baseline understanding, which can be used to make several educated guesses about algorithms to use and features to consider.

2. Applying a variety of shallow (Naive Bayes, Decision Tree, Logistic Regression, etc.) and deep (RNN, CNN, GAN, etc.) learning algorithms and techniques, where applicable, to fit our datasets with high accuracy.

3. Developing predictive models from the data to be applied in future business decisions and applications.

2. Project Approach and Management Plan

2.1. Overview

The primary deliverables of this project are the mid-term and final reports. These reports shall include updates of the progress of the project in achieving the three objectives outlined in Section 1.3, above. The deliverables will include all appropriate figures, charts, and visualizations of the exploratory analysis, shall detail the methods of data processing and analysis, and shall provide clear evaluation of the results.

Deliverables:

a. *Mid-Term Progress Report:* This progress report will provide an update on the status of each of the three objectives, and demonstrate any results. This report will also consider the time-line for delivery of the Final Report, and the remaining steps and schedule necessary for completion. This deliverable should include a preliminary version or working copy of the Final Report.

b. *Final Report:* This report will detail all of the actions taken in carrying out the project, and shall follow an IEEE conference or journal formatting. Specifically, this report should include Introduction, Background, Data Description, Evaluation, and Conclusion sections. *Note: These headers are subject to revision as necessary.*

Dataset:

Amazon Customer Reviews: The acquired dataset contains some 130 million customer review instances. Dataset features include: marketplace, customer_id, review_id, product_id, product_parent, product_title, product_category, star_rating, helpful_votes, total_votes, vine, verified_purchase, review_headline, review_body, and review_date. The data is collected into 46 ".tsv" (tab-separated value) files.

2.2. Schedule

Based on the overview provided, two major reports are to be delivered. The first, the Mid-Term Progress Report, shall be be delivered by midnight on *November 8, 2018*. The second, the Final Report, will be delivered by midnight on *December 6, 2018*.

2.3. Qualifications

This project team has been recently formed combining expertise in machine learning and marketing analytics. In addition, the team has some minor experience in the use of the Python programming language to carry out Data Mining tasks, as well as more extensive experience applying TensorFlow for machine learning classification tasks. While this project will present several novel challenges, we believe this team is able to combine their diverse skills and knowledge to accomplish the outlined tasks.

3. Evaluation of Success

The success criteria of this project will be based on two major factors: (i)The content and novelty of the two primary deliverables (Mid-Term Progress Report and Final Report) and (ii)The completion of the three Objectives outlined above. Note that the completion of the objectives heavily influence the content of the Final Report and, to a lesser extent the Mid-Term Progress Report. The rubric for assessing these reports can be found in the *COSC757_Assignment Evaluation Rubric.pdf* provided by the professor, unless any superseding rubric is provided.