IDS 575 Statistical Models and Methods

Term project- First Report

Project Title: Automated product score generation based on sentiment analysis on customer reviews

Contents:

INTRODUCTION TO DATASET

Amazon product data is a collection of customer reviews for products belonging to specific categories carried by the e-commerce site. The collection has 24 product groups with a total of 142.8 million reviews collected over the timespan of May 1996 - July 2014.

Data Source:	Julian McAuley, UCSD
Data Description:	The data contains reviewerID, productID(ASIN), reviewerName, helpfulness, reviewText, overallScore, summary, unixReviewTime and reviewTime

For computational ease, the analysis is focused on the product group-Clothing, Shoes and Jewelry which has 278,677 reviews in total.

OBJECTIVE 1

The primary objective of the project is to process the rich text from customer reviews corpus and convert them into features that can be inputs to any learning algorithm. Pre-processing of raw text includes the following steps:

- **Tokenizing** sentences to obtain individual tokens/words.
- **Normalizing case** of the raw text (usually converting everything to lowercase).
- **Stemming** tokens to retain only the root form of the words.
- **N-gram** construction to retain contextual meaning of the text.
- **Stopwords** removal to eliminate filler words from becoming features.
- **Punctuation** removal to have only words as features.

As a result of the above mentioned pre-processing steps, the corpus of raw user generated reviews is now converted into a bag of words (also termed as vocabulary in text-mining terminology) that can be used in generating features for our statistical model.

FEATURE GENERATION METHOD:

Assigning weights to each term in our vocabulary gives us information about the content of the document, sentiment associated to each document that can be learnt to provide predictions or classification.

The TF-IDF matrix (Term Frequency-Inverse Document Frequency) is a good measure to represent the importance of terms in our vocabulary. Each review in the corpus is considered as a document and all the terms in our vocabulary are the features that have a numeric value associated to it.

Now the data is ready to be passed through a learning algorithm to learn the relationship between our features and response variable.

OBJECTIVE 2

The variable- overall score is customer's rating for a product which is nothing but a numeric representation of the customer's review or opinion of the product. Since the universal rating system is on the scale of 1-5, the caveat to using the customer rating values as a metric to make any purchase or business decision is the subjectivity involved in choosing the rating. In other words, 2 customers who have given positive reviews to a product can choose either 4 or 5 (assuming 4 & 5 used for higher rating) as their rating which is completely subjected to the individual's level of satisfaction. This subjectivity affects the aggregated rating values of each product and decreases their reliability while making decisions.

The central idea of this analysis is to standardize and automate product score generation which will increase the reliability of the value. This is based on the sentiment captured from each customer review.

Hypothesis: Sentiment analysis of customer reviews is a direct measure of the overall product score.

Method: Build a logistic regression model that identifies sentiment associated to each review and represents the results on a scale of 1-5 (similar to overall score). The sentiment score is compared against the original overall score to measure the model performance and parameter tuning.

The critical assumption for the central idea to hold good is that the overall score in training data is a standardized representation of the scores with very minimal subjectivity.

EXPECTED CHALLENGES

- Threshold values to bin the sentiment scores is a trial and error process that will need domain knowledge (product group knowledge in our case).
- TF-IDF values may not be sufficient as features to our model. Other feature generation methods may give completely contrasting results.

- R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016
- J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015