Austin and Cedric

Big Data Report unformatted

**Introduction**

Why does one customer choose to purchase one item over another? If two items are truly substitutes (i.e., they perform the same function), there must be some defining features or aspects that lead to purchase of specific products. These are common questions asked across all industries and even across microeconomic academic circles.

Traditionally, when trying to identify the causal factors of a consumer’s purchase two methods are typically utilized:

1. Experiments with a control group and treatment group
2. Surveying customers followed by a conjoint analysis

Both of which come with strong disadvantages. Traditional experiments are incredibly costly and time consuming outside of the software industry (i.e., AB testing), whereas survey analysis introduces strong bias where a participant may answer in the form that they believe the surveyor wants them to.

Our research develops a complete methodology for understanding the relative importance of a product’s aspects within the retail sector. Our method leverages the unstructured text in product reviews to identify aspects, clustering algorithms to select the common aspect phrases, sentiment analysis to determine the user’s opinion towards the aspect, and then uses a probabilistic ranking model to evaluate each feature’s importance. We apply our model across a wide variety of unrelated product categories to demonstrate the generalizability of our model across the retail sector. Specifically, our model leverages the following pipeline:

[[insert diagram of pipeline]]

The result is that for each product, we will have identified the key product aspects and their relative importance. Since we identify aspects at the category level, we then are able to easily compare aspects and aspect importance over a wide range of products.

We must acknowledge that this is just the first step in causal analysis to understanding consumer reasons for purchases, but we must impart the two key advantages over traditional approaches to gathering such data:

1. The analysis time after model development is extremely quick, representing a significant improvement compared to experiments or surveys
2. Potential surveyee bias has effectively been removed by leveraging open-ended unstructured reviews volunteered by purchasers of the product without knowledge of this research

**Relevant Work**

Due to the multi-layered approach of our work, there exists multiple bodies of literature that overlap with our analysis including opinion mining, sentiment analysis, and aspect ranking. We provide a basic overview of some of the work from each of these that have influenced our analysis.

Opinion mining:

* [[product aspect ranking and its applications]]
* [[the use of language models to extract aspects – found in the above paper]]
* [[one more]]

Sentiment analysis

* [[Bing Liu’s seminal sentiment analysis paper]]
* [[ULMFiT paper by Howard and Ruder]]

Aspect Ranking

* [[product aspect ranking and its applications]]
* [[one more]]

**Data**

Our project leverages millions of unstructured Amazon product reviews developed by Julian McCauley and his team for [[insert the two citations]]. The Amazon Product reviews data consists of two primary datasets: 1) Product Reviews and 2) Metadata. Both of which are utilized.

After De-duplicating, the Product Reviews dataset consists of 88 million observations. Within our analysis we utilize the following columns:

* ASIN: Product ID number
* Reviews: Unstructured review text
* Overall: number of stars (i.e., score) given to the product for the review

Additionally, the Metadata dataset consists of approximately 9 million products. From this we utilize the following columns:

* ASIN: Product ID number
* Categories: Complete list of categories associated with the product

Note, that due to time constraints, we have limited the number of reviews for each top-level category to 1 million. We incorporate [[33]] top-level categories in our analysis.

**Methodology**

[[Insert more detailed diagram of the process]]