**Big Data Project – Report Outline**

1. Introduction
   1. What current issue we are trying to solve
   2. Basic overview of how we plan to solve this
   3. What the expected results / applications are
2. Relevant work
   1. Bing Liu’s Sentiment Analysis paper
   2. Product Aspect and Ranking paper
   3. One or two others
3. Data
   1. Amazon Data
   2. Reviews
   3. Metadata
4. Our Approach
   1. Figure 1: Flow chart of the entire process
   2. Aspect extraction
      1. How we extracted the aspects
         1. Dependency parsing and grammar structure
      2. Implicit vs. Explicit features
   3. Clustering
      1. Methodology for clustering
      2. Removal of outliers
   4. Sentiment analysis
      1. Methodology and model
      2. How we split sentences
         1. Essentially differentiating between how the model was trained and what was actually evaluated
      3. [[Anything else?]]
   5. Generating Weight vectors
      1. Methodology
         1. [[since we are basing this off another paper, I do not think that we need to go too in-depth here and can just reference the paper]]
      2. Figure 2: math equations [[we don’t need the whole proof]]
5. Results
   1. Sample of Aspects extracted
   2. Sample of Clusters and representative word identification
   3. Sentiment Analysis results
      1. Table 1: scores of sentiment analysis model
   4. Final results
      1. Table 2: overview of the data (number of reviews and aspects by top-level categories)
      2. Table 3: top 10 features for the five main categories
      3. Table 4: lowest 10 features for the five main categories
      4. Table 5: top 5 electronics categories with top 5 features from each
      5. Table 6: comparison – same 5 electronics categories, comparing a sample of 10 features across them
   5. Summary
      1. While we only showed a small snippet, our results are for hundreds of thousands of products. We developed an unsupervised generalized approach for arbitrary aspect extraction and valuation. [[a paragraph along these lines]]
6. Future work and considerations
   1. Improvements to be made:
      1. Better aspect extraction
         1. Discerning between descriptors (implicit) and noun chunks (explicit) features
      2. Better clustering
         1. DBSCAN which allows for outlier detection during clustering
      3. Removal of outliers from cluster
         1. Understanding that aspect extraction is not perfect, we remove outliers within each cluster to ensure that aspects are not overrepresented
      4. [[anything on your end?]]
   2. Future work
      1. Causal aspects for purchasing or not-purchasing
      2. Copyright infringement
      3. Marketing
      4. [[Thoughts?]]
7. References