INTRODUCTION
Consumers consider customer reviews posted on ecommerce websites before making a purchase decision. Unfortunately for popular websites and popular products often there are far too many reviews that make it difficult for a prospective buyer to read through all reviews and make a decision. This poster demonstrates the application of text mining techniques to collect, group and summarize positive and negative opinions on a product. For the purpose of this poster, I have developed a tool using ASP.NET to extract (1674) customer reviews for Kindle Fire HD 7" from Amazon.com website as of October 31, 2012. Then, by importing this excel worksheet as a SAS dataset, text mining can be performed to summarize customer comments by grouping related reviews into clusters. The Text Parsing, Text Filter, Text Topic and Text Cluster nodes are used and outputs from every node are discussed. Decision Tree is used to predict the rating a customer would provide based on his reviews. Sentiment Analysis is later performed to develop a model to classify positive and negative reviews.

DATA COLLECTION
- The reviews available on Amazon.com are in a user readable format and can be navigated from page to page by clicking the page index buttons.
- To analyze these reviews, one must convert them to a SAS data set. So a type of web crawler software tool was developed using ASP.NET technology to parse the html source code of every review page and extract each user review and export it to an excel (.xls) file.
- This reusable tool would take the review page URL as the input and would crawl across all review pages for a particular product. The tool filters all irrelevant html tags, empty spaces, special characters and other symbols from the reviews page and extracts just the customer reviews from all pages.

DATA CLEANING
The excel data is imported into SAS using the File Import node. The variables used can be found by selecting exported data property of the file import node. In the Text Filter node, the default option of Check Spelling has been changed from "No" to "Yes". The Ignore Parts of Speech option is set to "Abbr, Aux, Conj, Det". The most frequently used words (eg. Do, the, is, and etc.) and the least frequently words used are filtered by preparing a filter data set and exported using the Text Filter node.

This synonym data set is fed into the Text Parsing node so that these words are parsed in the Parsing node. In the properties panel of Text Parsing node, the Different Parts of Speech option is set to No. A custom Stop List is prepared and fed into the Text Filter node. Once the synonym data set and the custom stop list data sets are fed into the Text Parsing node, the Text Filter node now automatically filters out the most frequently used and least frequently used words based on the synonym list and cleans up the data accordingly.

CONCEPT LINKS
- By clicking the Interactive Filter Viewer in the Text Filter node, we can find how strongly a term is associated with another term using Concept Links. Concept link diagram of Kindle Fire HD 7" shows that it is mostly associated with great, buy, book, read, love.
- Most of them are positive comments made by the reviewer.
- Further refining of term great shows it is related to terms such as easy, sound, price suggesting people have made comments on the ease of use of Kindle Fire HD 7".

TOPIC ANALYSIS
The Text Topic node performs cluster analysis to group the documents and summarizes the collection by identifying "topics". These topics are generated by combining words that are interesting to analysts.

CLUSTER ANALYSIS
The Text Cluster node produces 8 clusters with each cluster talking about different aspects of Amazon Kindle Fire HD 7". Cluster 2 talks about the camera, sound, speed and battery life of the product.
- Cluster 1 talks about some good points about the product.
- Cluster 4 mainly talks about the problems associated with Kindle Fire HD 7" and has a number of dissatisfied customers.
We filter the SAS data set by classifying observations into positive and negative using the individual star ratings. We have classified reviews with 4 and 5 stars as positive reviews and 1, 2, 3 stars as negative reviews. Since SAS Sentiment Analysis Studio accepts only text file input, we convert the SAS data sets generated for Text Mining into individual text files using a customized SAS Macro code. Since the sample contains around 1100 positive reviews and around 600 negative reviews, we have taken around 600 positive reviews for training and validation purpose and saved the remaining 500 positive reviews for testing purposes. We now develop a Statistical training model using Model based sentiment analysis approach with 70% of the data for the training and 30% for validation. Best Mode was chosen before training the statistical model.

### Conclusion

In our research, we analyzed reviews of one product from one web site and observed some interesting patterns in the general comments made by the customers. Reviews of the same product can also be compared across many sites such as eBay, BestBuy and find how the reviews may be different among these sites. While Text Miner is a good first step, sentiment mining using SAS Sentiment Studio can perhaps provide deeper insights into these reviews. These results will be useful to both product manufacturers for product improvement and customers for making purchase decisions.

Nagarajan, Duraidhayalu, and Chakraborty