Investigating the Impact of Amazon Kindle Fire HD 7” on Amazon.com Consumers Using SAS® Text Miner and SAS® Sentiment Analysis

Srihari Nagarajan, SAS Institute, NC, USA
Hari Harasudhan Duraidhayalu, Kavi Associates, IL, USA
Dr. Goutam Chakraborty, Oklahoma State University, Stillwater, OK, USA

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.

Figure 1. Data preparation process for Text Mining

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 the Speech option is set to “Abbr, Aux, Conj, Det”. The most frequently used (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 was created based on frequencies of terms extracted by applying the Parsing node one time and then subjectively incorporating terms that are not suitable for this analysis. A custom dictionary dataset was also 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.

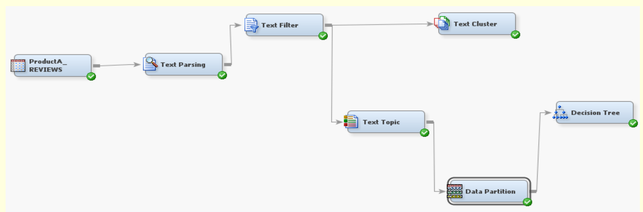


Figure 2. Text Parsing and Filter nodes for Data Cleaning

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".

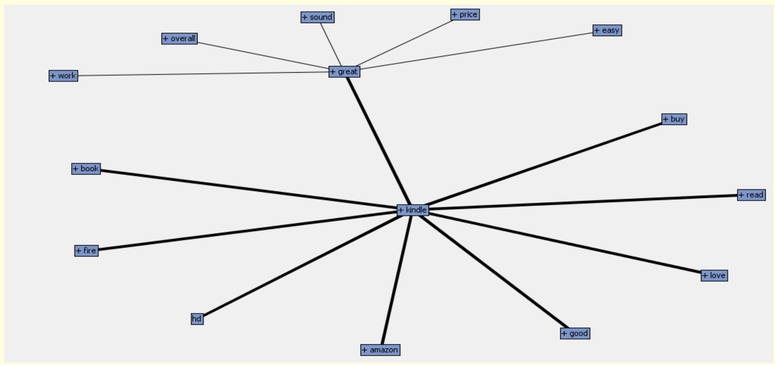


Figure 3. Concept Link diagram showing relationship between terms

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.

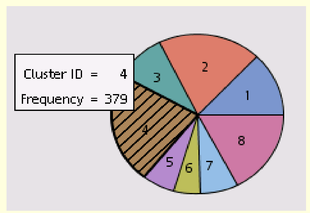
Topic	Number of Terms	# Docs
+battery,+life,battery life,+hour,+charge	144	218
+book,+read,+game,+music,+play	158	266
+book,+read,+text,speech,+device	232	256
+browser,+page,+web,silk,+carousel	226	244
+button,+power,+volume,+device,+power button	212	222
+camera,+skype,+picture,+face,+original	172	232
+card,+credit,credit card,free,us	129	164
+case,+cover,+find,+easy,+charger	248	226
+charger,+charge,usb,+wall,+plug	131	224
+device,+pay,+remove,+offer,+carousel	208	263
+download,+file,+device,+video,+book	233	235
+flash,+support,adobe,+video,+work	160	240
+great,+sound,+tablet,+quality,+picture	177	280
+ipad,+apple,+device,+small,+good	172	254
+kfhd,kf,+device,+generation,+camera	217	211
+love,+old,+new,+fire,+year	167	277
+nexus,+google,+tablet,+store,+amazon	155	228
+order,+amazon,+device,+kindle,+new	204	269
+original,+original fire,+new,+fire,hd	144	227
+play,+game,+device,+review,+people	220	235
+prime,free,+amazon,+content,+tablet	216	254
+watch,+movie,+tv,+video,+hdmi	183	253
+wifi,+connection,+fast,+original,+device	258	250
audio,wi-fi,+dolby,dual-band,+special	112	190

Figure 4. Topics grouped by terms and docs from the Text Topic output

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 *books*, *movies, games* and *ease of use.*
- Cluster 4 mainly talks about the problems associated with Kindle Fire HD 7" and has a number of dissatisfied customers.



Cluster ID	Descriptive Terms		Frequency	Percentage
1	+love kindle books +fire +great great hd +read games movies +tv +price +sound +easy +new	...	216	13%
2	+battery +life far +size reading +easy +music great +great speakers +book movies amazing +device +fast	...	323	19%
3	+support +content +flash flash stars +browser +web +overall +work +amazon +tv +play +watch out +download	...	162	10%
4	disappointed +software up +experience +new +problem +work +time returning +browser +download flash +pay +support +return	...	379	23%
5	+google +nexus play +store +tablet +play +amazon +browser games +experience +flash speakers +good +device out	...	100	6%
6	+return returned returning +ipad +support own flash +charge +download +big usb +amazon +new +tablet +nexus	...	84	5%
7	+buy bought buying kindle hd +fire +purchase +love +happy books +fast +recommend best movies +need	...	124	7%
8	+button +charge +charger +pay +power charging usb +volume +big +overall stars +find one +purchase speakers	...	286	17%

Figure 4. Terms grouped as clusters with their frequencies listed

PREDICTION OF INDIVIDUAL RATING

The individual star rating of the product can be predicted based on the terms used in the review using a decision tree model. The topics are fed as input to the decision tree after a 70:30 data partition for training and validation purposes. The model's ASE is close to 13%. The table below shows the important topics used in prediction of star rating in the decreasing order of importance.

Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	TextTopic_raw7	+return,+device,+buy,+amazon,+ipad	1	1.00000	1.00000	1.00000
2	TextTopic_raw17	+buy,+nexus,+annoy,+disappoint,+don	1	0.94744	0.79311	0.83711
3	TextTopic_raw19	+disappoint,+amazon,+find,+ipad,+pay	1	0.75896	0.53738	0.70804
4	TextTopic_raw8	+card,credit card,+credit,+buy,us	1	0.69994	0.40373	0.57681
5	TextTopic_raw1	+love,+buy,+amazon,+fire,+device	1	0.46440	0.28098	0.60504
6	TextTopic_raw12	+support,+browser,+flash,flash,+support	1	0.44538	0.62009	1.39229
7	TextTopic_raw11	+pay,+remove,+amazon,+advertisement,+device	1	0.40866	0.00000	0.00000

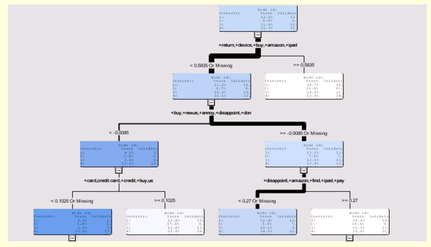
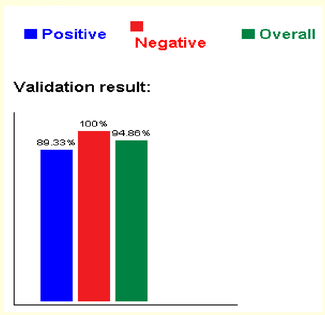
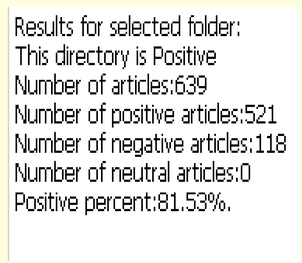


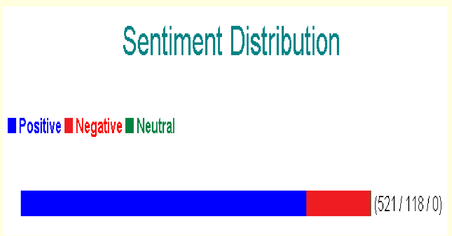
Figure 5. Important variables for prediction of rating listed with partial decision tree diagram

SENTIMENT ANALYSIS

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.







The Sentiment Analysis model has an overall accuracy of 94.86% in the validation data. It has a positive accuracy of 89.33% in the validation data. The same model when applied on testing data reveals that the 81.53% of the positive sentiments are classified as positive. So we can conclude that the model has an accuracy of 81.53% in the test data.

The SAS macro code used for splitting the reviews data set is given below.

```
/*WORK1: START******/
dm "log; clear; output; clear;";
%macro createtxt(numvars);
%do i=1 %to &numvars;
   data _null_;
   obsnum=&i;
   length text $2500.;
      set source POINT=obsnum;
      file &&name&i;
      text=compbl(strip(tranwrd(comment,"","")));
      put text;
      STOP;
   run;
%end;
%mend;
data _null_;
   set source nobs=count;
call symput("name"||left(_n_)," 'C:\DestinationFolder"||Id||".txt' ");
  if _n_=1 then call symput("numvars", trim(left(put(count, best.))));
run;
%createtxt(&numvars);

/*WORK1:END*/
```

Note: To get the customized web crawler tool code written in ASP.NET kindly contact srihari.nagarajan@sas.com

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