University of Connecticut

Logo, company name

Description automatically generated

OPIM 5671 - Data Mining and Business Intelligence

Project: Amazon Electronics Reviews

Professor – Sudip Bhattacharjee

By

TEAM – 6

Akshay Kumar Pydi

Amulya Boinapalli

Divya Sai Teja Pamarthi

Jessica Tattersall

Mayank Patidar

Contents

[Executive Summary 3](#_Toc80457005)

[Business Problem 4](#_Toc80457006)

[Methodology 4](#_Toc80457007)

[Sample 4](#_Toc80457008)

[Explore 5](#_Toc80457009)

[Modify 6](#_Toc80457010)

[Model 7](#_Toc80457011)

[*Modeling – Review Body* 7](#_Toc80457012)

[*Modeling – Review Title* 16](#_Toc80457013)

[Assessment 18](#_Toc80457014)

[*Review Body* 18](#_Toc80457015)

[*Review Title* 19](#_Toc80457016)

[Conclusions 20](#_Toc80457017)

[Recommendations 21](#_Toc80457018)

# Executive Summary

E-commerce giant Amazon wants to analyze the reviews given by customers to predict sentiment, determine authenticity, and understand key issues in the electronics department. We took a dataset from Kaggle and randomly sampled 50000 records from the original dataset. We preprocessed the data by removing duplicates, blank entries, and irrelevant entries. The dataset contains 90% positive ratings and 10 % negative ratings.

Words like good, product, quality, sound, phone, battery, charge, not, can’t. don’t, and won’t played a key role to determine sentiments. We tried different term weights, clusters, and text topics, and models. To determine the correlation and authenticity between the review body and review title, we applied the same modeling on both columns. Regression models with custom text topic nodes yielded the best results.

The built model will help the e-commerce client understand their sales performance and customer sentiment. The client can then tailor the product according to reviews to further improve satisfaction and ultimately increase sales. It is also important to display the reviews with the highest document weights which most accurately summarize the product and customer satisfaction. This will then help attract future customers. The e-commerce client must remain attentive to these product reviews and sentiment or risk damaging their product sales and reputation. Many sellers today even respond to customer reviews in the hopes of attaining customers they would have otherwise lost. There is a lot of responsibility by the seller to respond to customers to not only retain or improve sales but also build a solid reputation that they can leverage to deploy new products more easily. Ultimately this is a great example of the power of communicating meaningful metadata.

# Business Problem

Electronics in Amazon is one of the highest revenue-generating departments. With the increase in the number of electronic products, the authenticity of the product is determined by the ratings given by previously purchased customers. With the increase in the number of reviews coming daily and the dependency of sales on these reviews, the e-commerce client felt that there should be a check on the genuineness of the ratings and review comments. For example, in a few scenarios, the rating was given as 5 out of 5 but the review title and body have negative comments with many cons on a product. We will try to create a text mining model to transform unstructured review comments into a structured format to identify meaningful patterns and new insights.

# Methodology

## Sample

This dataset is taken from Kaggle and contains over 400k rows and 6 columns. For our analysis, we selected 50,000 rows randomly where we made sure to remove duplicate records and avoided the missing value records. The initial rows of the dataset are shown in the screenshot below:

Graphical user interface, text, application

Description automatically generated

Dataset Details

* The first column is the Name of the reviewer or customer. This column is insignificant for our modeling part.
* The "title" column depicts a short description of the review that an individual has posted. For example, in row 7 it says, "Very good headphone", from this comment we can predict that the overall review posted might be a positive response from the customer or user.
* The 3rd column is "Review body" which contains the full review posted by the user. We have used this column for EDA and text clustering purposes.
* Review Rating is the actual rating out of 5 that the user has given, based on his experience with the product.
* The sentiment column is basically either Positive or Negative, based on the rating given by the user. If the rating is either 4 or 5 the sentiment is positive, else it is negative.
* The Sentiment\_binary column is nothing but a binary assignment for the sentiment column, positive as 1 and negative as 0.

## Explore

We started with removing duplicates, blank entries, and irrelevant entries from the review body column. After cleaning the data, we have picked approximately 50,000 rows randomly. We found that approximately 40,000+ rows have a positive sentiment which means the review rating is either 4 or 5 out of 5. The remaining 9500+ rows have a negative sentiment with a rating of either 1,2 or 3. This can be verified in the pie charts below.

|  |  |
| --- | --- |
| Chart, pie chart  Description automatically generated | Chart, pie chart  Description automatically generated |

In the review body column, the most occurred words were good, product, quality, sound, phone, battery, charge, etc. From the word cloud, we made our analysis that most of the electronic products are cellular phones, headphones, cameras, or some battery-consuming gadgets.

Text

Description automatically generated with medium confidence

Word Cloud for Complete Dataset

## Modify

Sentiment\_Binary is a new binary column created from the original column Sentiment. For the initial sampling part of our dataset, we tried various combinations for the Data set allocation of the imported data. The data partition that we have used for training, validation, and test are 60,25,15 respectively.

Table

Description automatically generated

Data Partition

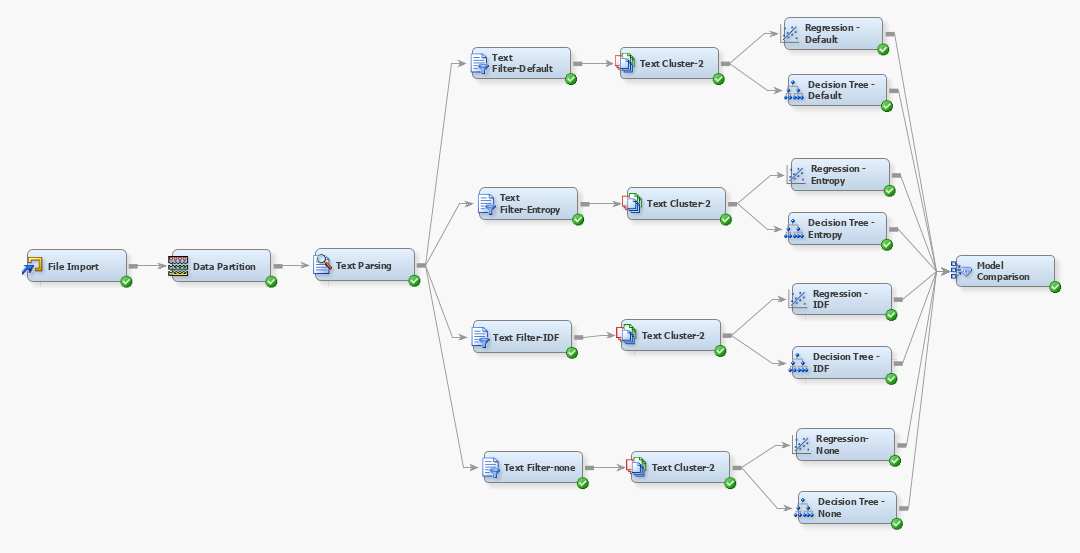
We started our model by rejecting all the columns except Review body (text role) and sentiment\_binary (target variable). To compare the results from this model, we changed our text variable to title and ran the same models for predicting the correct sentiment.

|  |  |
| --- | --- |
| *Review body as the text variable* | *Title as the text variable* |
|  |  |

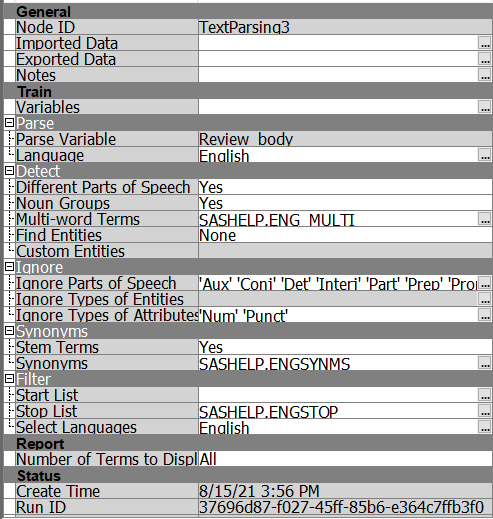
## Model

*Modeling – Review Body***:** Initially we have started with an INPUT = Review Body, Target = Sentiment\_Binary

**Analysis for Text Parser and Text Filter parameters:** Our Initial analysis was to figure out the suitable Text Parser and Text Filter configurations that would yield an optimal result.

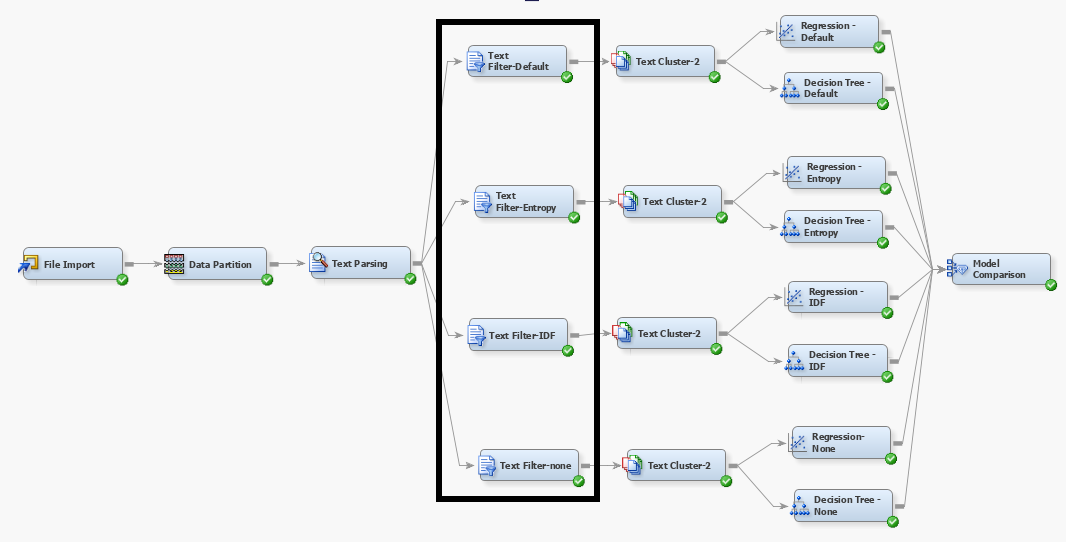


**Text Parsing:** We have used the Text Parsing node to extract the terms of the review body and then updated the stop lists to include the negative contractions words like not, can’t, wouldn’t, shouldn’t and others as these constitute the major terms in the text mining analysis. Rest all the parameters well set as the default settings as shown below.

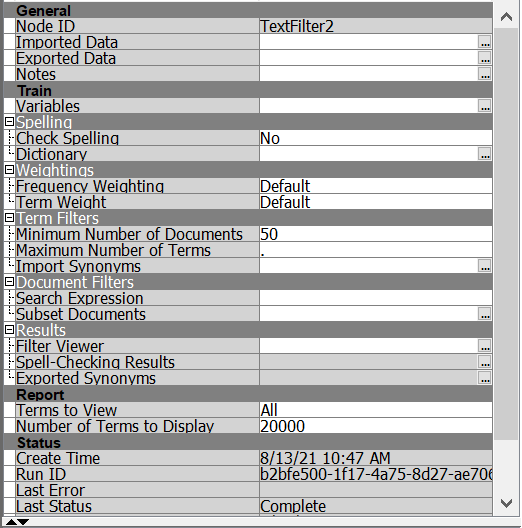


Text Parsing Node Parameters

**Text Filter:** We have attached multiple text filters to the text parsing node with the Frequency weight set as default because it reduces the effect of high-frequency words on the text mining process. The Term weight setting was changed in each node, and we added 4 nodes having Term weight as Default (mutual information), Entropy, Inverse Document Frequency, and None. The Minimum number of terms is set as 50.



Text Filters with various Term Weights and Frequency weight = Log



Text Filter Properties Panel

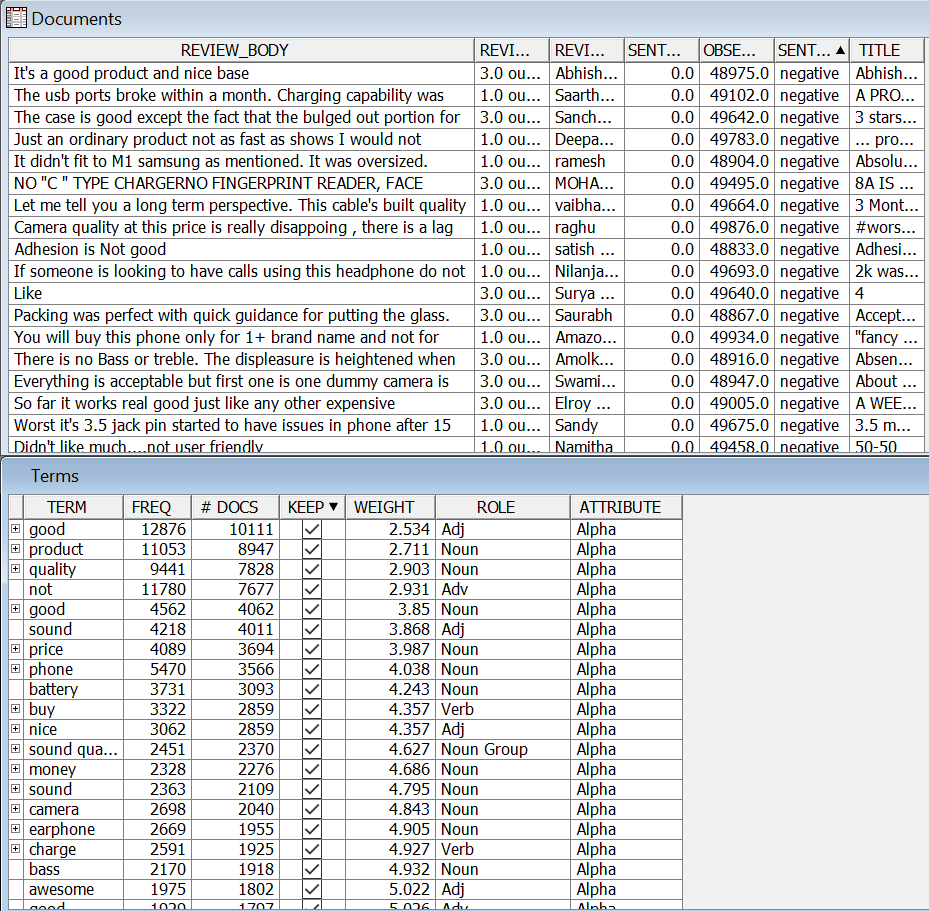
We have then attached regression nodes and Decision tree modes to all 4 types of Text filters and then compared the results to find out the best performing model. Below are the results:

Table

Description automatically generated

Assessment for Best text filter properties

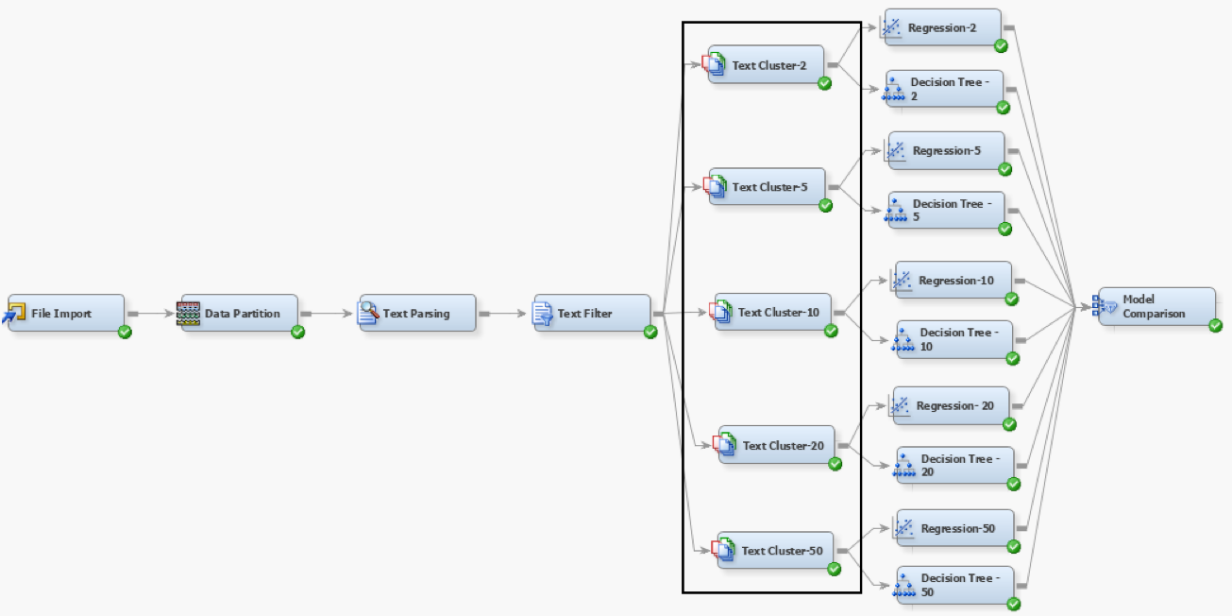
Hence, we can conclude that the Regression model with Text filter Term weight as Inverse Document Frequency and frequency weighting of log performs better compared to all other options. It has a maximum ROC of 0.9.



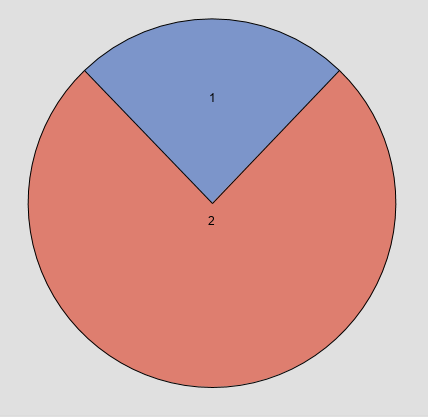
Interactive Filter Viewer Results

When we check the filter viewer of IDF Text Filter, we see that the Terms good, product, quality are the most occurred words and have a high weight in this dataset, thereby playing an important role in the prediction of the Sentiment of each document.

**Text Cluster:** To perform clustering analysis on each document to the sentiment, we have tried various options of some clusters as shown below. All the clusters are then connected to the respective Regression and Decision Tree nodes after which we have used the Model comparison node to compare the results. We have observed that Text Cluster 2 with SVD Resolution as Low and SVD Dimensions set as 100 would yield optimal results consisting of 49 SVD variables.

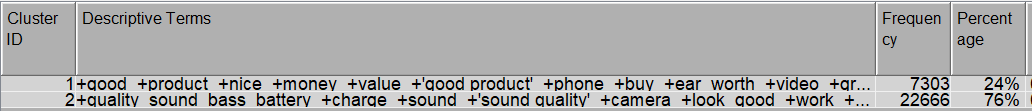


Text Cluster Analysis



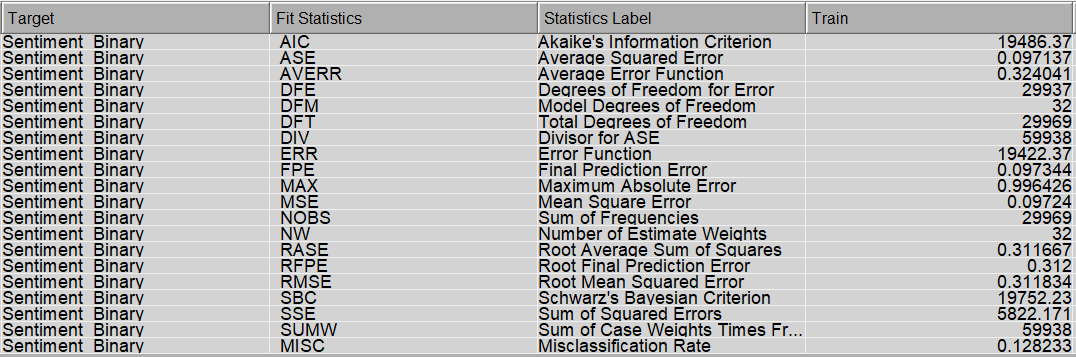
Cluster Frequencies

Cluster Frequencies and Cluster Details show the details of 2 clusters that were formed. We can see that Cluster 1 has 7303 documents included and cluster 2 has 22666 documents. This categorization helps in grouping out the Sentiment binary variable more effectively.



Cluster Details

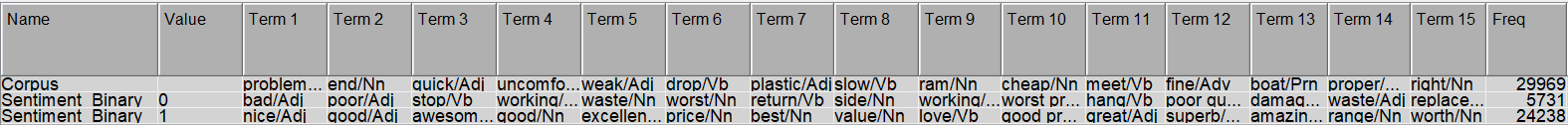
Below are the details of the optimal model that we obtained among various clusters that were analyzed. We can see that the Text Cluster 2 related Regression model has a maximum accuracy of 87.2% and is the best model to be considered.



Text Cluster 2 Regression Model Results

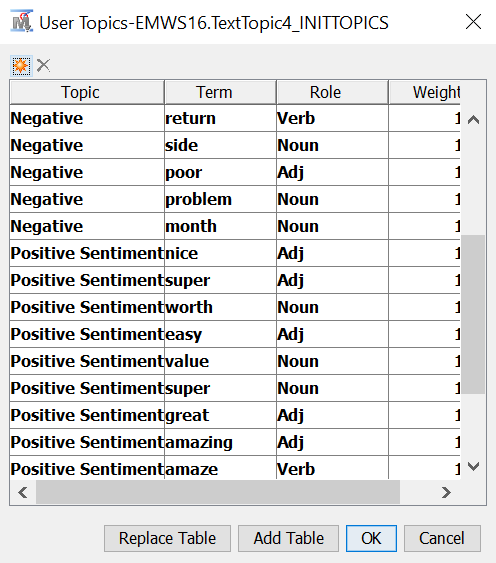
**Text Topic:** We have performed 3 types of Analysis on the Text Topic Nodes which we have described in detail below:

1. **Text Topic Analysis 1:** As part of this analysis, we have used a Text Profile with 15 terms to find out all the possible terms that will be formed in each sentiment. Below are the details of the 15 most frequently occurring terms for each sentiment.

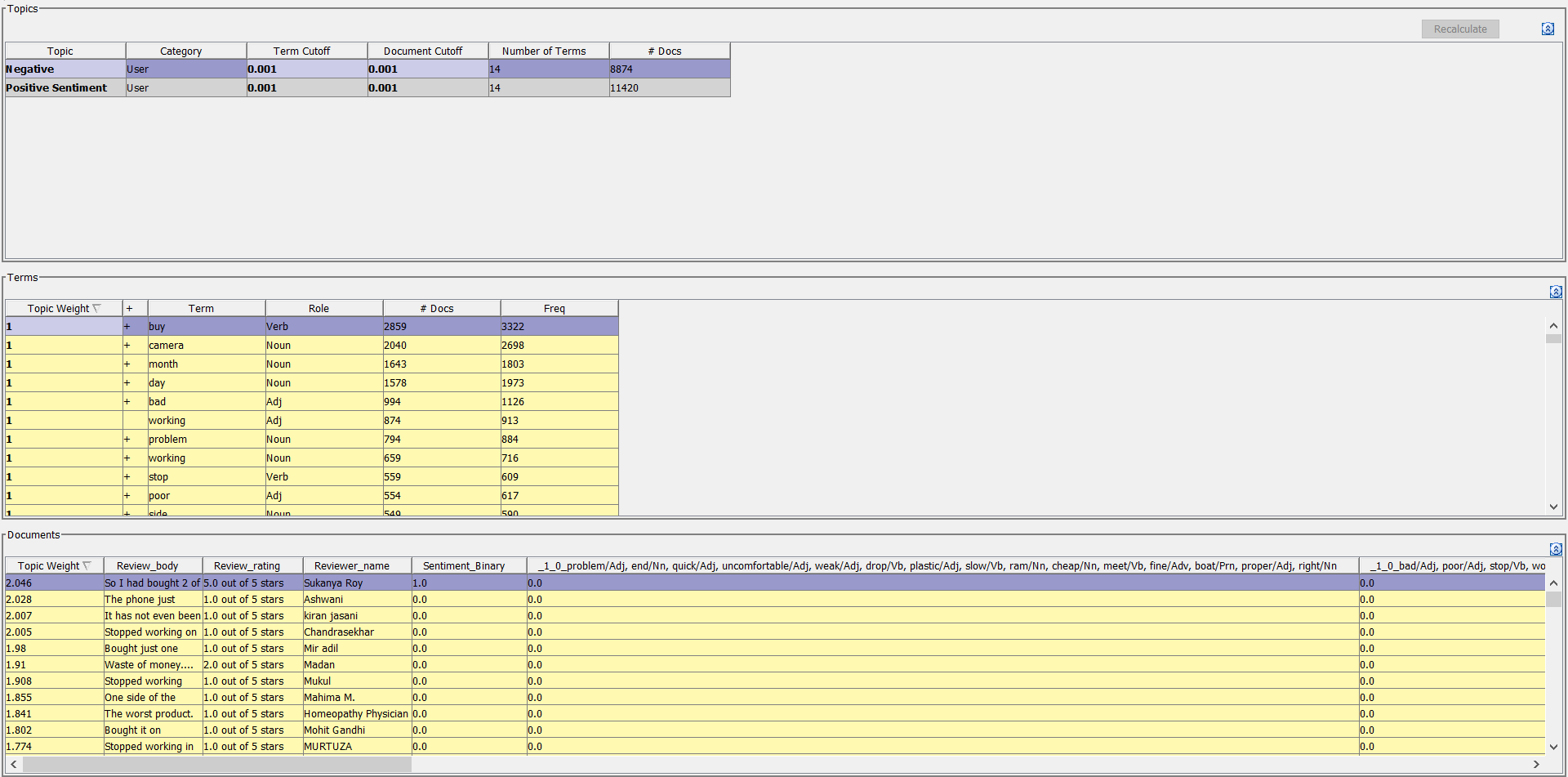


Text Profile Results

The Text topics are then added to the User topic for each of the sentiments and named accordingly as shown below. With this, we were able to group various documents under Positive and Negative sentiments as we can see in the below Interactive filter viewer results.

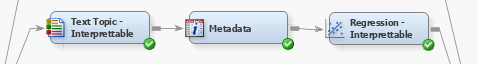


User Topic Addition



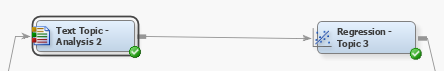
Interactive Filter Viewer Results

1. **Text Topic Analysis 2:** This is known as the Interpretable model where we have used a Text topic and default setting are assigned to it and then the Text Cluster Probability and Text Cluster variables are set as input parameters using metadata node for the Regression Model.

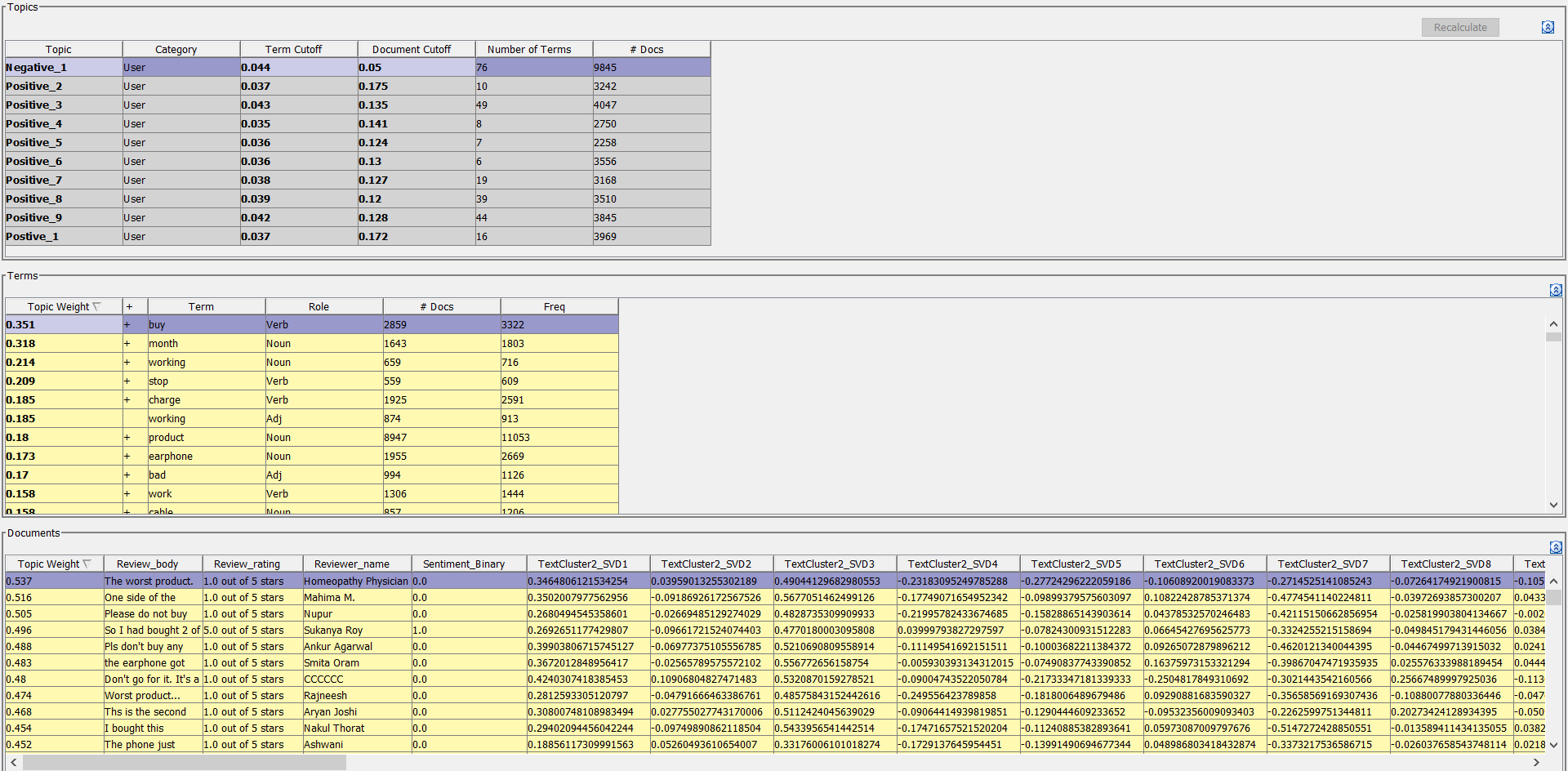


Text Topic Analysis 2

1. **Text Topic Analysis 3:** For this model, we have taken 10 Multi terms in Text topics and then assigned the Term cut-off and document cutoff accordingly based on the usage of terms and their values. 1 topic has negative sentiment, and the remaining have a positive sentiment. The Interactive Filter Viewer shows the 10 Text topics that were formed, and we can see that one Text Topic is finally concluded as Negative while the rest are Positive Sentiment.



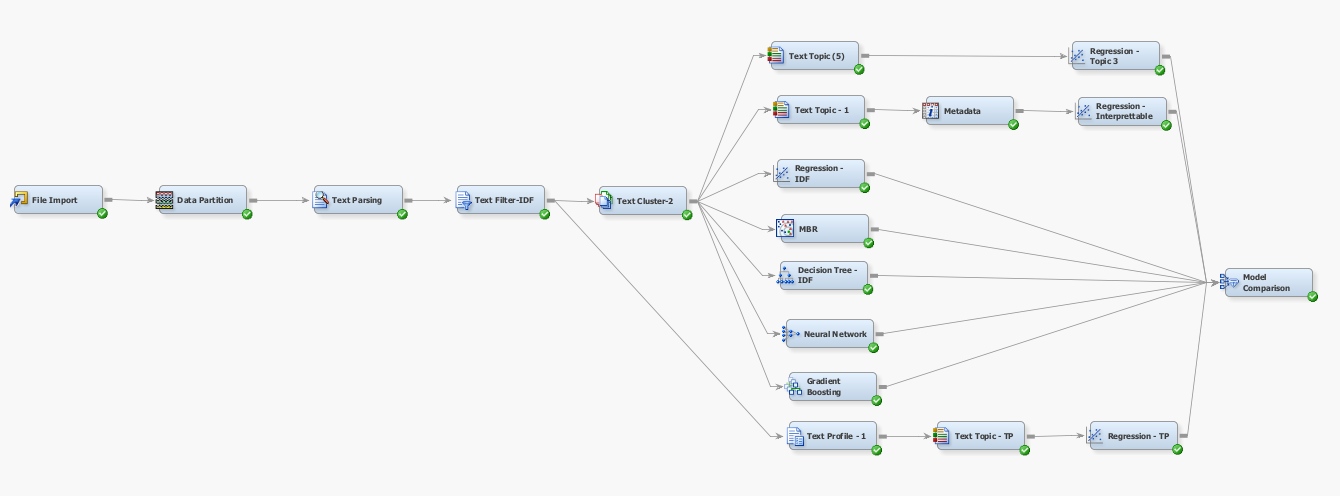
Text Topic Analysis 3



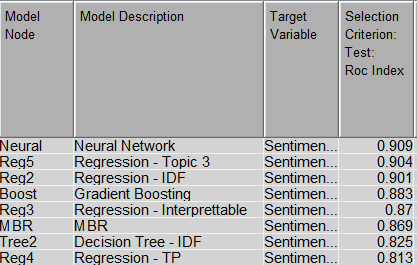
Interactive Filter Viewer Results

**Model Comparison – Review Body:**

Below is the final diagram used for modeling and then for comparison of the results. We have used various classification models like Regression Model, Gradient Boosting model, Neural Network, MBR. We see that the Neural network model has higher ROC followed by Text topic 3 analysis and Regression IDF model.



Final Modelling Diagram



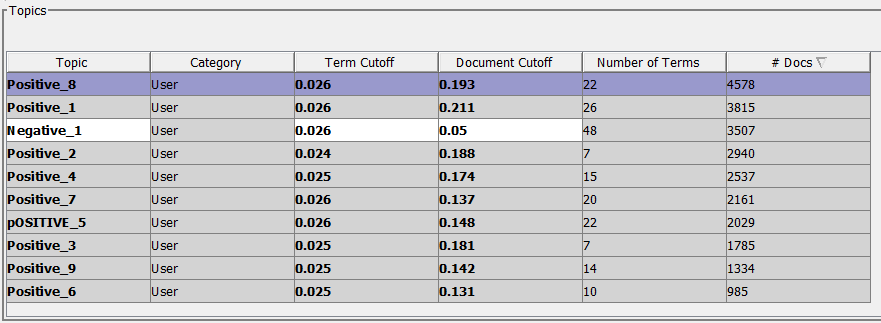
Model Comparison Results

*Modeling – Review Title***:** Now, we will be using the INPUT = Review Title, Target = Sentiment\_Binary

To determine the correlation and authenticity of the reviews, we decided to apply the same modeling strategy which was applied on the Review Body column. We were able to find out that the Text Parsing, Text Filter, and Text Cluster nodes results obtained are like what we have got for Review Body. Hence, we then focused on the Text Topic node and Text Profile node.

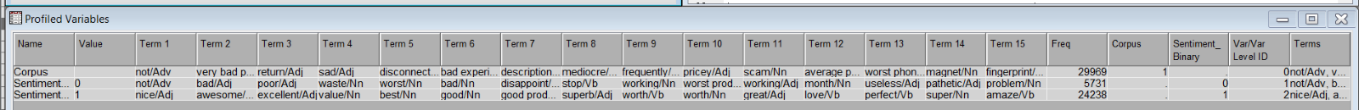
**Text Topic (5):** We used some multi-term topics as 10. This generated 9 positive topics and 1 negative topic. ‘not’ played a key role in the review title to determine negative sentiment with a weightage of 0.883. To cover a greater number of documents the document cutoff has been decreased by 0.05. We manually checked a few words and gave the term and document cutoff accordingly. We made sure that the negative sentiment-related words that we found during the exploration are clearly identified and added to the Text topics term cutoffs accordingly.

Similarly, positive sentiment words that we found during exploration are added to the positive text topics which are 9 in number in our case.



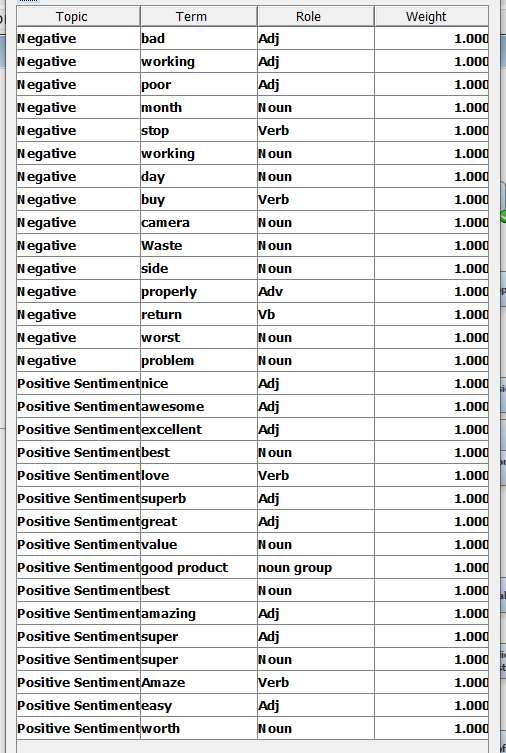
Text Topics

**Text Profiler - 1:** We selected a maximum number of terms as 15 to find out all the possible terms that will be formed in each sentiment. The below terms are the most frequently occurring 15 terms for each sentiment.



Text Profile Results

The Text topics are then added to the User topics for each of the sentiments and named accordingly as shown below.

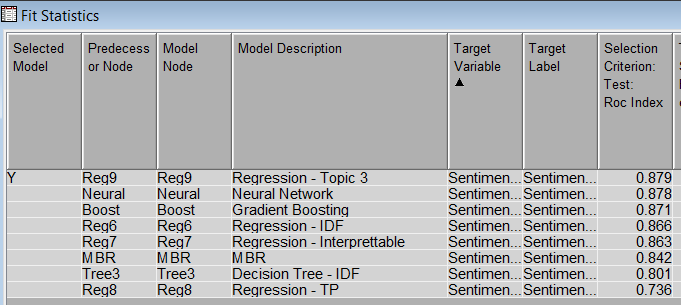


User Topic Addition

Both the Text Topic and Text Profiler are connected to the regression model.

**Model Comparison – Review Title**

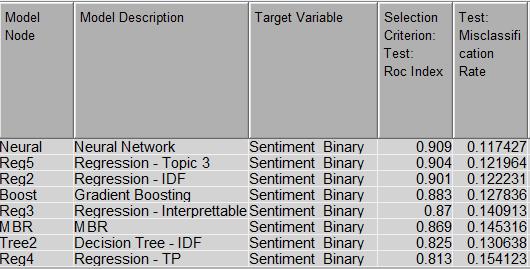
The following shows a model comparison of all the model's results. We have used models like the same as Review body analysis. We see that the Regression Node of Text Topic 3 has a maximum ROC of 0.879 and can be easily interpretable.



Model Comparison Results – Review Title

Assessment

### *Review Body*



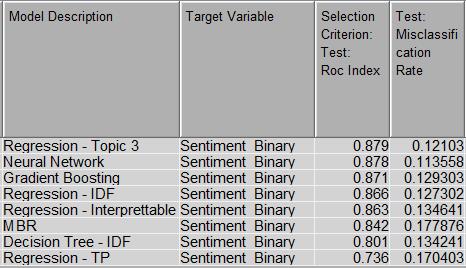
Review Body Results

The ROC value is highest for the Neural Network model with a value of 0.909, closely followed by the Regression model run on Text Topic 3 with a ROC value of 0.904. The neural network model is a black-box model and is thus difficult to interpret. If the model is difficult to interpret, the analysis obtained from it cannot be used for business purposes. Hence, even though the neural network model behaves slightly better than the regression model, it is rejected.

The Regression model is interpretable and has the next best ROC value of 0.901 with an accuracy of 88% and is thus recommended as the best model to use for Amazon reviews.

### *Review Title*

The ROC value of the Regression model run on Text Topic 3 has the highest ROC value of 0.879 which is only slightly better than the ROC value of 0.878 for the Neural Network model. Since the Regression model has the best performance and is also interpretable, we recommend that Amazon use this model to analyze review titles. This model has an accuracy of 87.9%.



Model Comparison results – Review Title

# Conclusions

* The most found words in the reviews are good, product, quality, sound, and battery. Mostly these reviews talk about sound quality, camera, and battery life. From this, we can assume that the categories of electronic products that are most reviewed are mobile phones, earphones, speakers, and cameras.

Text

Description automatically generated with medium confidence

Negative Review Words

Graphical user interface, text

Description automatically generated

Positive Review Words

* We found that the Title of the reviews and Review Body has similar key descriptive terms and almost always has the same sentiment. Thus, looking at the titles of reviews would be sufficient to understand the common sentiment associated with that product if the user does not want to read all the reviews.
* There is a misclassification rate of 11% for the review body and 12% for the review title. Negative sentiment has the most misclassification because multiple rows with a rating of 3 have mixed sentiments. There is a need for the classification of rating 3 words to avoid the mixed review comments.

# Recommendations

Below are few recommendations which can be effectively followed by Amazon to enhance the productivity and sales of the venture:

* Amazon can use the words associated strongly with positive sentiment to design a dropdown for every product that allows the user to filter reviews by features that could be advantageous for that product. Similarly, they can use the words strongly associated with negative sentiment to design a dropdown that allows users to filter reviews that could be disadvantageous for that product.
* Amazon can design another dropdown using the words associated with features like battery life and sound quality, which were found common in many of the reviews. This could allow the user to filter reviews by the features they are looking for in a product.
* Having a Title selection dropdown to look at only the best, moderate, or worst reviews based on the user selection. This would let a user read the reviews by sentiment and get a better picture of the common sentiment for the product.

These dropdowns make it convenient for users to find the reviews with the content they are looking for, so they don’t have to read all the reviews.

* It is best to display the reviews of positive and negative sentiment that have the highest document weights with the product images to summarize that product.