OPIM 5671 – DATA MINING AND BUSINESS INTELLIGENCE

AMAZON ELECTRONICS REVIEW

Mayank Patidar

Problem Statement

- 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 there should be a check on the genuineness of the ratings and review comments.
- For example: in 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.

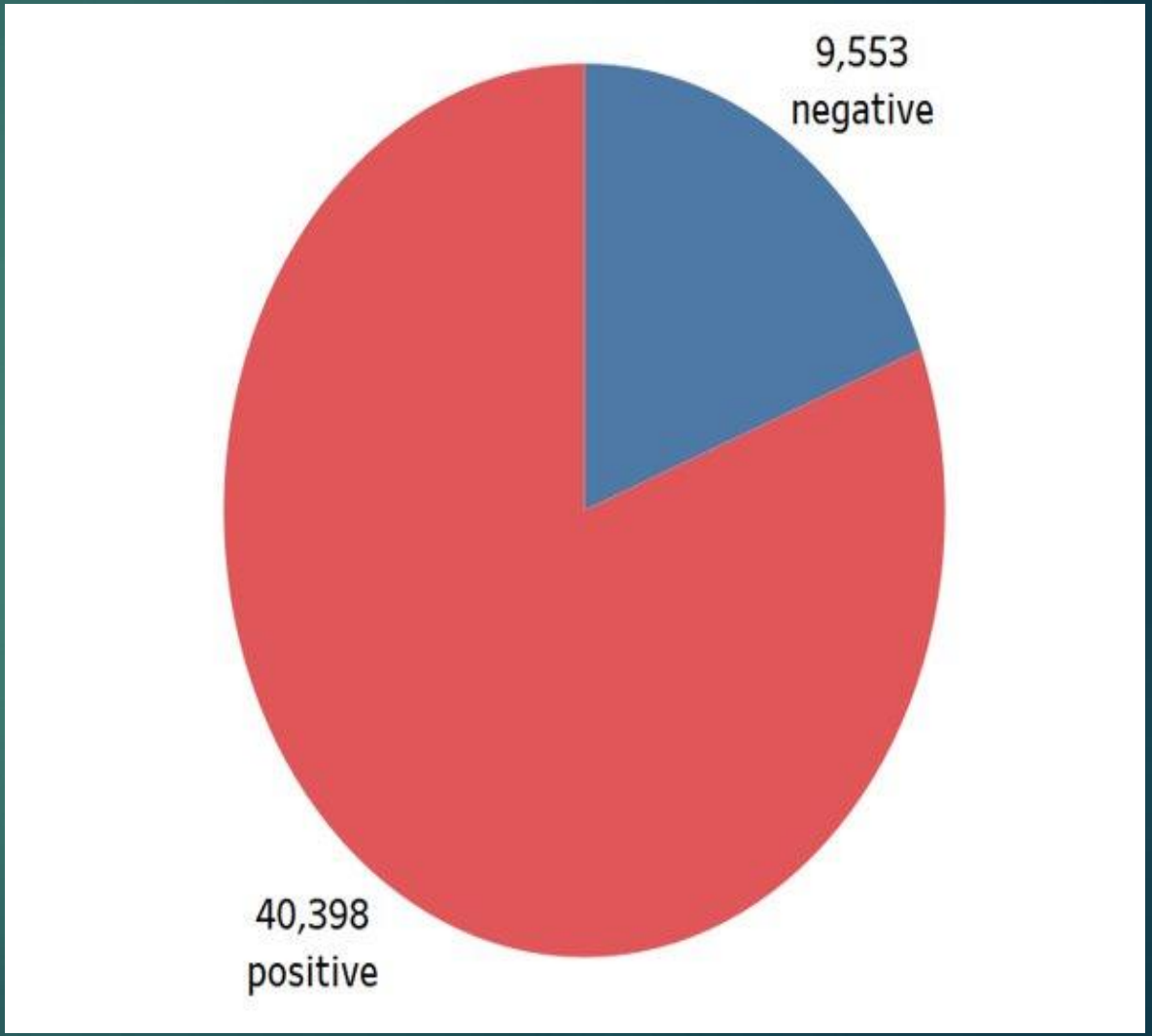
Sample

- We have 6 columns in our dataset
- **Reviewer name** is the person who has reviewed the electronic product
- **Title** is the title for the actual review
- **Review body** contains the whole comment of reviewer
- **Review rating** is rating out of 5
- **Sentiment** is either positive or negative
- **Sentiment Binary** is the binary value of sentiment, 0 stands for negative and 1 for positive

Reviewer name	title	Review body	Review rating	sentiment	Sentiment_Binary
Aaditya lucky	Okay	It's okay I would sa	4.0 out of 5 stars	positive	1
Aaditya P.	Very bad quality product	From the day it has	1.0 out of 5 stars	negative	0
Aaditya Sahu	Quality	Awesome device at	4.0 out of 5 stars	positive	1
AADITYA SANKALP	Good in bass	Good in bass, good	4.0 out of 5 stars	positive	1
aaeeyna and vivek	Solved somethings for me	I have to place my A	5.0 out of 5 stars	positive	1
Aahan Ratnaparkhi	Flop model in one plus..👎	The product is not u	1.0 out of 5 stars	negative	0
Aahil Nabi Khan	Very good headphone	Sound and bass is e	4.0 out of 5 stars	positive	1

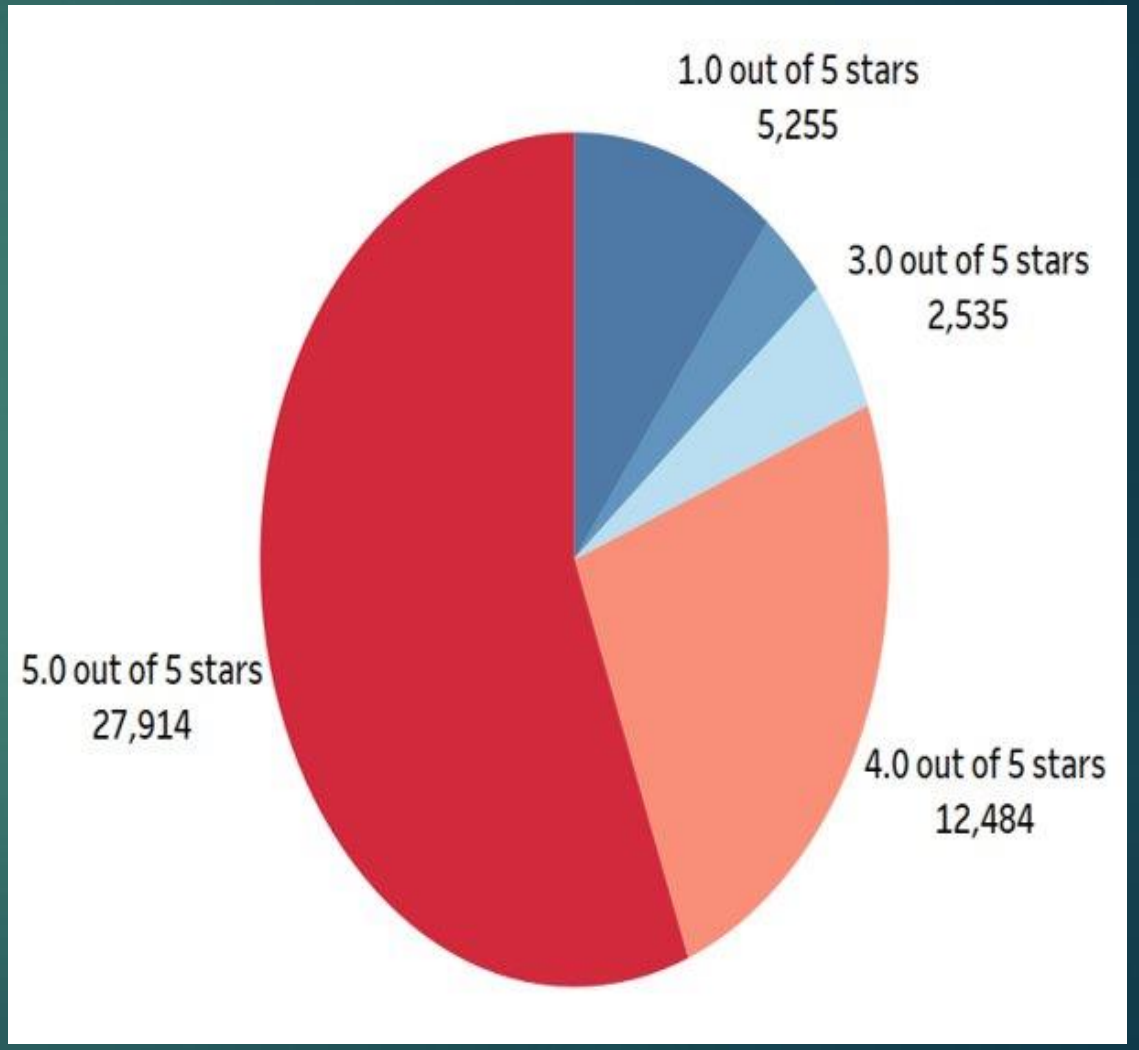
Data Exploration

- There were approximately 50,000 records in the dataset.
- The average length of reviews comes close to 230 characters, lengthier the review more helpful it is.
- Majority of the records (40,000+) were **Positive** sentiments.
- There were close to 9500+ rows with the sentiment as **Negative**.



Data Exploration...

- The ratings were given on a scale of 1 to 5. if the review rating.
- A positive review is either 4 or 5 and negative review is either 1,2 or 3.
- Majority of the records (40,000+) were either **4 or 5** out of 5.
- The negative sentiment group has almost 9500+ records and consists of review rating of **1,2 or 3**.

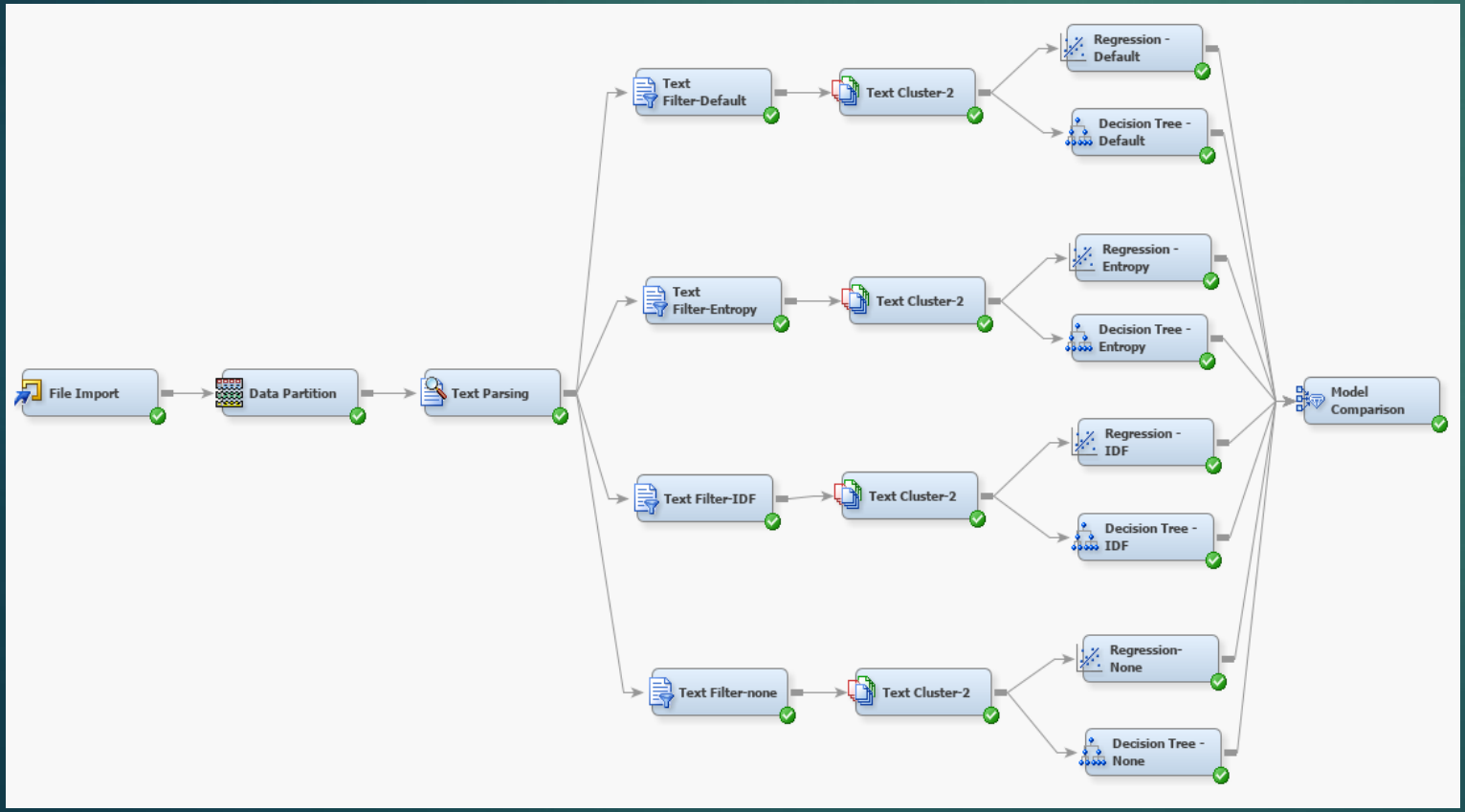


Modeling – Review Body/Review Title

Partition – 60:25:15

Text Parsing and Text Filter

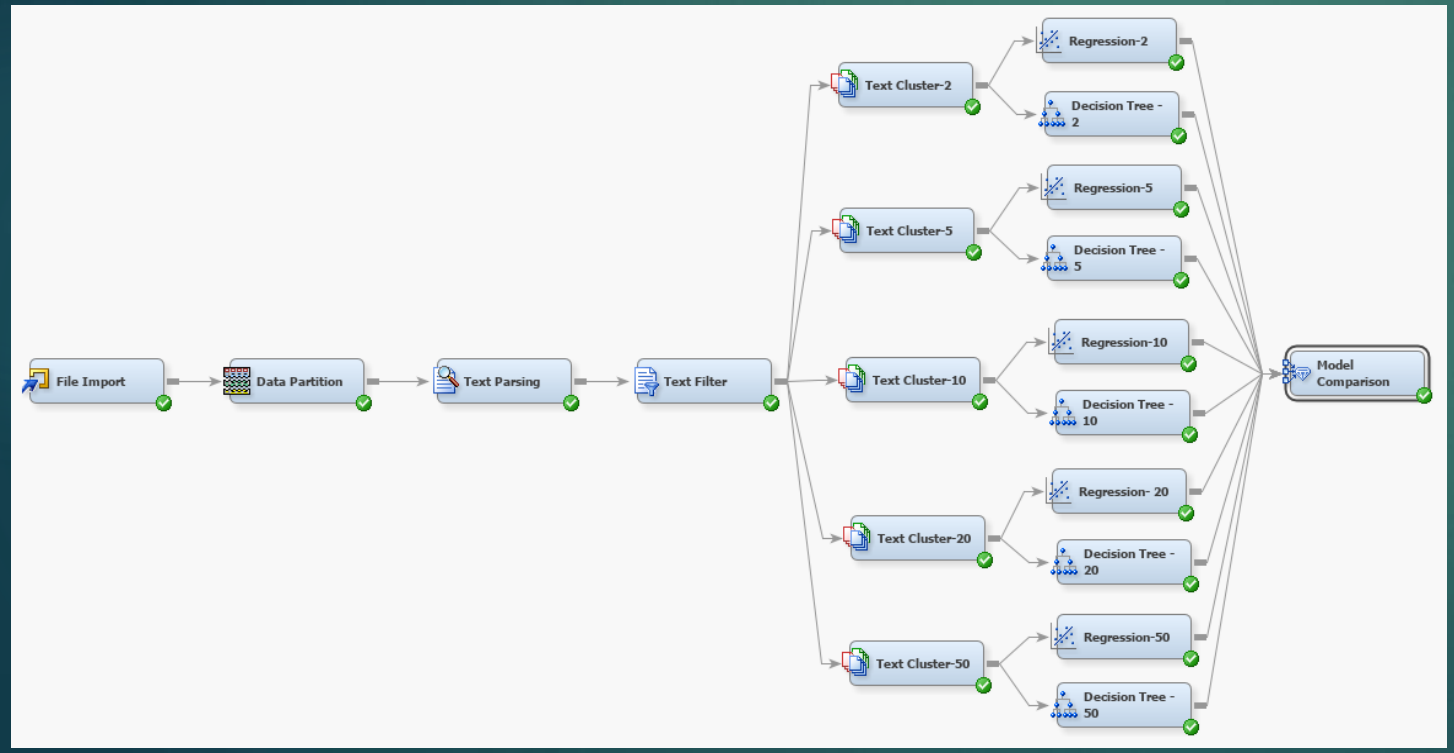
- Terms like not, can't, won't and other contraction words are removed from the stop list.
- Inverse Document Frequency Term weight yields the best results in terms of ROC.
- The minimum Number of terms in Document is validated and finally set as 50.

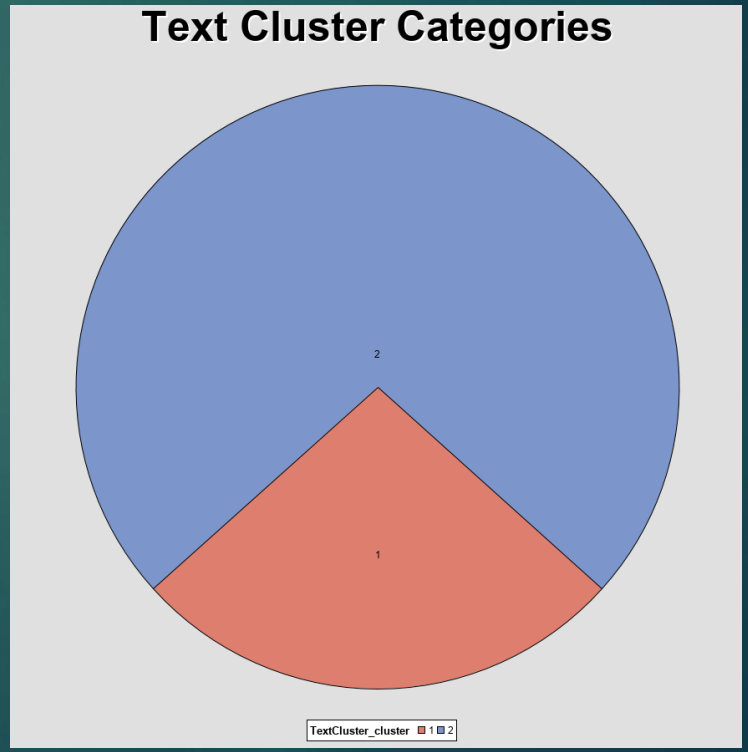


Model Description	Target Variable	Selection Criterion: Valid: Roc Index
Regression - IDF	Sentiment Binary	0.9
Regression - Entropy	Sentiment Binary	0.899
Regression-None	Sentiment Binary	0.858
Regression - Default	Sentiment Binary	0.851
Decision Tree - Default	Sentiment Binary	0.83
Decision Tree - Entropy	Sentiment Binary	0.823
Decision Tree - IDF	Sentiment Binary	0.822
Decision Tree - None	Sentiment Binary	0.786

Text Cluster

- We have tried for various clusters and could see out that the results remains consistent.
- Text Cluster 2, with SVD Resolution = Low and Number of SVD Dimensions = 100 has the better ROC value.
- Optimal Number of SVD variables created is 48





Text Topic

- 3 Different types of Analysis were done to make sure the grouping of Terms is done effectively.

Analysis 1

- Interpretable model having the TextCluster_Prob and TextCluster_n variables set as Input

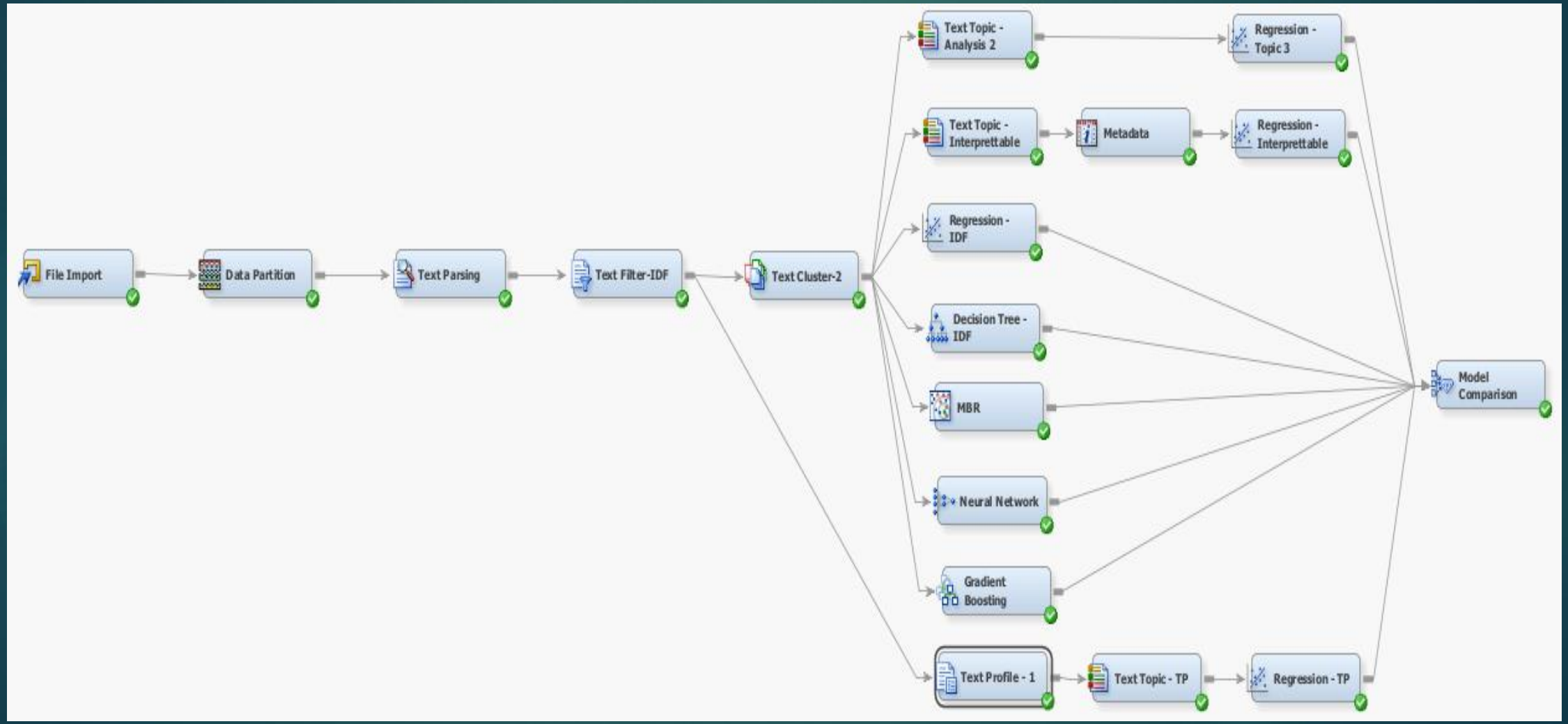
Analysis 2

- Assigning of the Term and Document Cutoff based on the word's presence in the document

Analysis 3

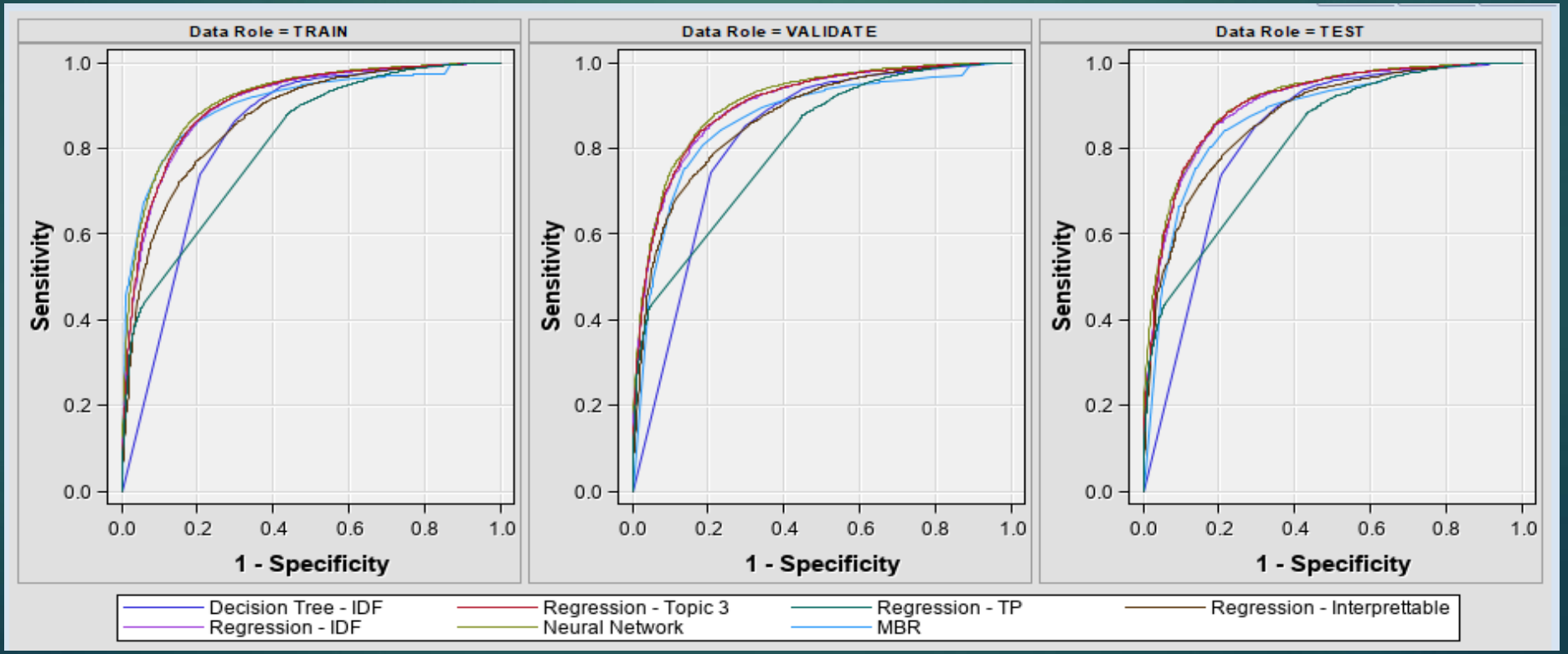
- Using Text Profile to cluster 15 most frequent terms for each category

Diagram


File Import
Data Partition
Text Parsing
Text Filter-IDF
Text Cluster-2
Text Topic - Analysis 2
Regression - Topic 3
Text Topic - Interprettable
Metadata
Regression - Interprettable
Regression - IDF
Decision Tree - IDF
MBR
Neural Network
Gradient Boosting
Text Profile - 1
Text Topic - TP
Regression - TP
Model Comparison

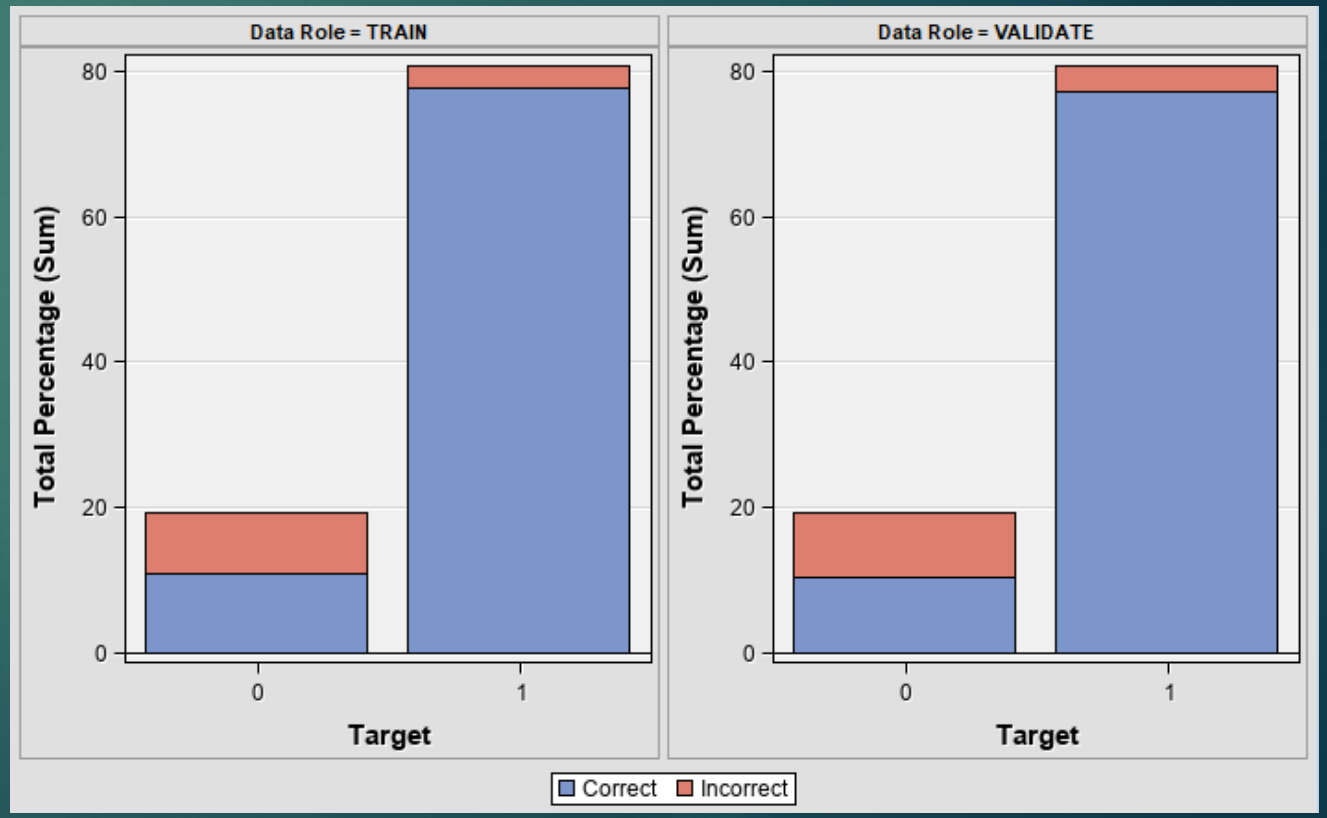
Top 3 Models – Review Body

Neural Network	Regression Text Topic	Regression Inverse Document Frequency
• ROC – 0.908	• ROC – 0.902	• ROC – 0.9



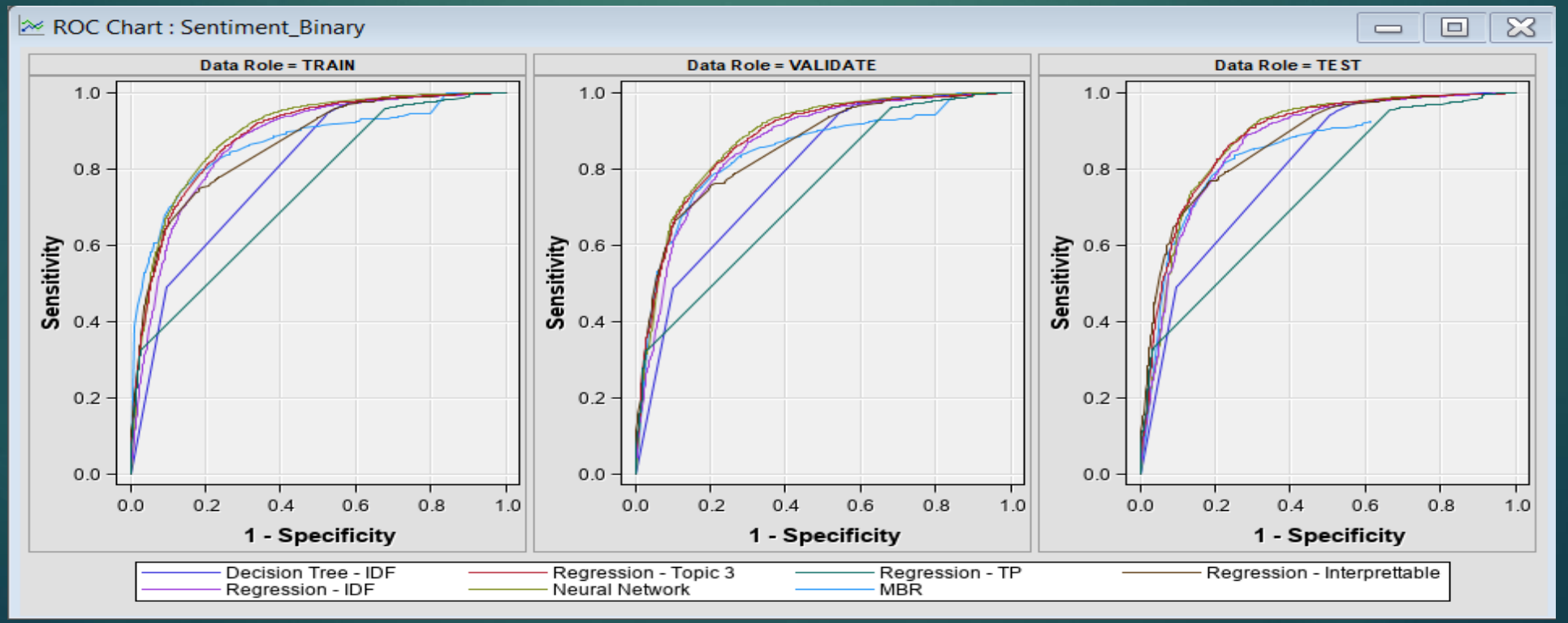
Best Model - Review Body

- Neural model is a black box model and hence cannot be a good interpretable model.
- Regression Model with Text Topic – Forward Selection and Validation Error Criteria
- Term Frequency = IDF
- Minimum Number of Documents = 50
- Optimal Number of Clusters = 2
- Optimal SVD variables = 48
- Text Topics – 10 Multi Term
- ROC = 0.902
- Misclassification Rate = 0.117



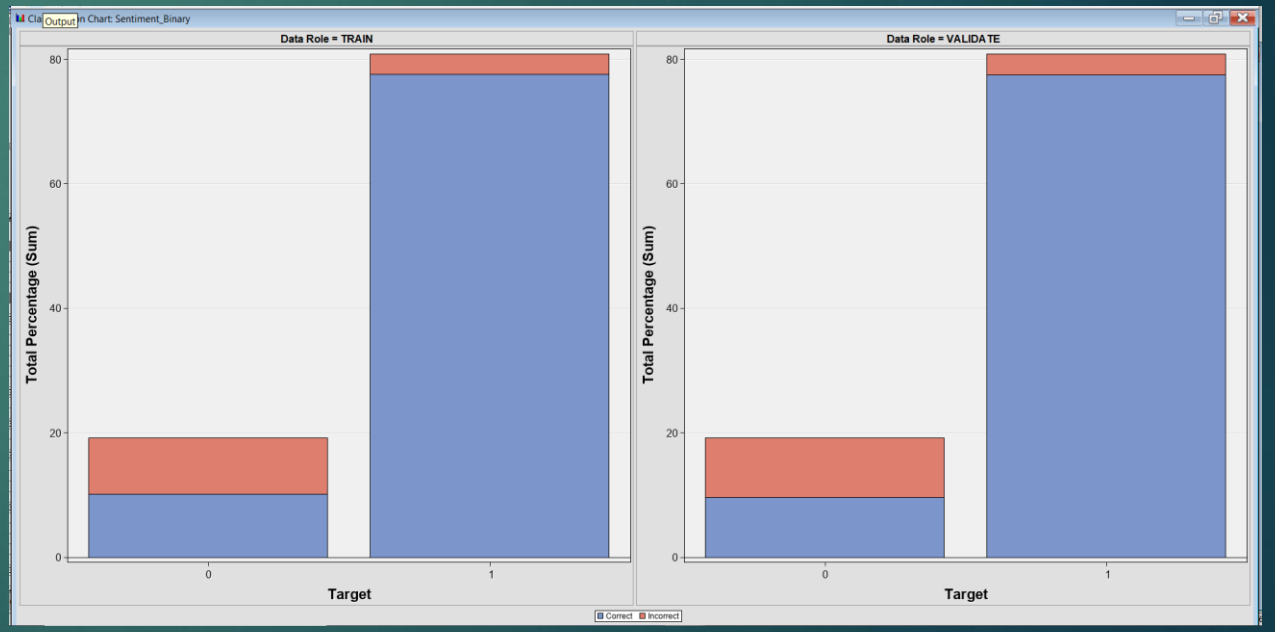
Top 3 Models – Review Title

Neural Network	Regression Topic 3	Gradient Boosting
• ROC – 0.88	• ROC – 0.875	• ROC – 0.863



Best Model – Review Title

- Neural model is a black box model and hence cannot be a good interpretable model.
- Regression Model – Forward Selection and Validation Error Criteria
- Term Frequency = IDF
- Minimum Number of Documents = 4
- Optimal Number of Clusters = 2
- Optimal SVD Variables = 43
- Text Topics – 10 Multi Term
- ROC = 0.875
- Misclassification Rate = 0.123



Key Findings

- The word cloud shows that most of the words in the review body column are good, product, quality, sound, battery.

- This shows that most of the products in the dataset were either cellular phones, speakers, headset, and cameras.

- Most of the comments are related to battery and quality of sound and camera for the products.

Negative Review Body Word Cloud

Positive Review Body Word Cloud

Key Findings..

Phones	• From the data, we know that most of the review comments pertain to the phones category where customers express their perspective on the pros and cons of the respective products.
Review Title (vs) Review Body	• We see that the terms of occurrence of Review body and Review title are similar and showing the same category of Positive or Negative Sentiment.
Rating 3	• There are 2535 rows having Rating as 3 which mostly have a mixed sentiment that involved users mentioning positive and negative review comments.
Negative Sentiment Misclassification	• There is more misclassification of the negative sentiment due to ratings of mark 3 and because of the mixed reviews by the users.

Recommendations

- Having a dropdown with most common good and negative reasons would improve user experience. Feature-based dropdowns can be implemented for enhanced model.
- Having a pre-defined "Title" selection dropdown(Best, Good, Moderate, Worst), so that users can directly trust the title if they are not willing to read the whole comment.
- Display the most useful reviews alongside the product images.
- Monitor the compensated-review processes that are based on verified purchases.
- Vet the users rigorously before giving free products for them to give honest reviews on.
- Monitor sellers who replace a product review with a high rated review of their own product, leading to mismatch of product and review.
- Manufacturers can incorporate and review product attributes when a product is launched and over time correct product issues, understand customer requirements, and maintain customer satisfaction.


thank you
danke
謝謝
gracias
merci
obrigado
dziękuję
bedankt
sukriya
kop khun krap
go raibh maith agat
teşekkür ederim
ngiyabonga
спасибо
Баярлалаа
рахмат
dank je
tapadh leat
mochchakkeram
terima kasih
grazie
arigatô
dakujem
takk
taiku
sagolun
hvala
mauruuru
faafetai lava
kiitos
dankie
dhanyavad
nandri
nanni
bayarlalaa
gràcie
köszönöm
enkosi
sobodi
dēkuji
mesi
didi madloba
kam sah hamnida
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
diolch
rahmat
rahmet
tanemirt
najis tuke
ありがとう
sulpáy
gràcies
gràcias ago
chnorakaloutioun
akun
dankon
aciù
vinaka
спасибі
blagodaram
mersi
kia ora
barka
welalin
tack
spas
misaotra
matondo
paldies
grazzi
mahalo
djiere dieuf
tau
дякую
mamnun
chokrane
murakoze
obrigada
asante
manana
tenki
хвала
trugarez
мерси
merce
shukriya
dhanyavadagalu