UNIVERSITY OF MALAYA

Milestone 4

Data Mining

WQD7005

Name : Sidratul Muntaha

Matric : Wqd180079/17199116

Name: Nur Nuraini Afiqah Binti Rohanip

Matric : Wqd180115/17198049

Project Title : Text Mining on customer feedbacks of different laptop brands

Milestone 4 objective :

Interpretation of data & Communication of Insights of data (Individual)

Introduction :

Data interpretation refers to the implementation of processes through which data is reviewed for the purpose of arriving at an informed conclusion. The interpretation of data assigns a meaning to the information analysed and determines its signification and implications. The interpretation of data is designed to help people make sense of data that has been collected, analysed and presented. Having a baseline method (or methods) for interpreting data will provide a structure and consistent foundation.

Work Flow:

1. Data Cleaning
2. Data Storing in Hive
3. Accessing data from Hive
4. Interpretation of Data

    i. Sentiment analysis on python
    ii. Text mining on SAS enterprise

Sentiment Analysis:

Referring to the python code in github link, the sentiment analysis is being processed and the extracted and refined dataset is being used to do text mining on SAS software.

Text mining on SAS enterprise:

1) We upload datasets on sas enterprise through sas studio and then converted it in sas.bdat format for later processes.

Figure 1: SAS Studio

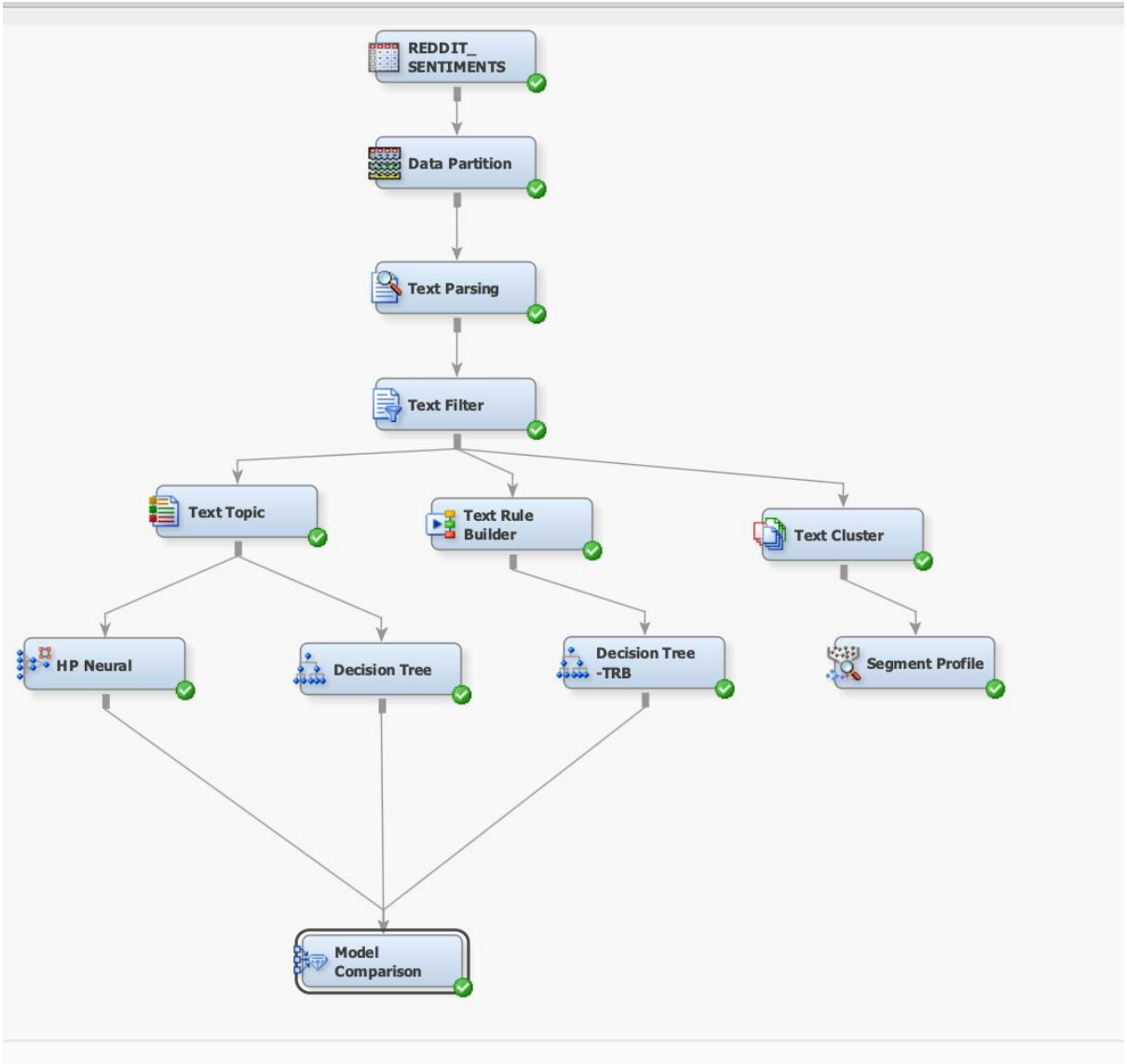


Figure 2: SAS Enterprise Miner Diagram

Tasks performed :

1. Identify the reddit data source with an **Input Data** node.
2. Partition the input data using the **Data Partition** node.
3. Parse the document collection using the **Text Parsing** node.
4. Reduce the total number of parsed terms using the **Text Filter** node.
5. Cluster documents using the **Text Cluster** node.
6. View the results.
7. Examine data segments using the **Segment Profile** node.
8. Applied ML algorithms and compared.

2) The **Data Partition** node enables to partition the input data into one of the following data sets:

- **Train** — used for preliminary model fitting. The analyst attempts to find the best model weights by using this data set.
- **Validation** — used to assess the adequacy of the model in the Model Comparison node. The validation data set is also used for model fine-tuning in the **Decision Tree** model node to create the best subtree.
- **Test** — used to obtain a final, unbiased estimate of the generalization error of the model.

We put 60% to train, and 20% to validation and 20% to test the data.

3) The **Text Parsing** node enables to parse a document collection in order to quantify information about the terms that are contained therein. **Text Parsing** node can be used with volumes of textual data such as e-mail messages, news articles, Web pages, research papers, and surveys.

We applied text parsing in order to quantify all the social media relevant words mainly used for commenting or giving feedback about laptops.

Term ▲	Role	Attribute	Freq	# Docs	Keep	Parent/ Child Status	Parent ID	Rank for Variable numdocs
+ account doe ...	Noun...	Alpha	1	1Y	+	16...	71...
+ account option...	Noun...	Alpha	1	1Y	+	10...	71...
+ accumulate ...		Alpha	1	1Y	+	12...	71...
+ acer ad ...	Noun...	Alpha	1	1Y	+	13...	71...
+ acer applicatio...	Noun...	Alpha	1	1Y	+	55...	71...
+ acer computer ...	Noun...	Alpha	2	2Y	+	15...	46...
+ acer et322qk s...	Noun...	Mixed	1	1Y	+	13	71...
+ acer forum ...	Noun...	Alpha	2	2Y	+	21...	46...
+ acer laptop ...	Noun...	Alpha	27	24Y	+	19...	10...
+ acer laptop wi...	Noun...	Alpha	1	1Y	+	490	71...
+ acer product ...	Noun...	Alpha	4	4Y	+	15...	30...
+ acer program ...	Noun...	Alpha	1	1Y	+	14...	71...
+ acer website ...	Noun...	Alpha	11	11Y	+	11...	16...
+ achieve ...		Alpha	8	8Y	+	54...	20...
+ acknowledge		Alpha	4	3Y	+	71	37

Figure 3: Text Cluster

4) The **Text Cluster** node clusters documents into disjointed sets of documents and reports on the descriptive terms for those clusters. The following is the cluster frequency chart.

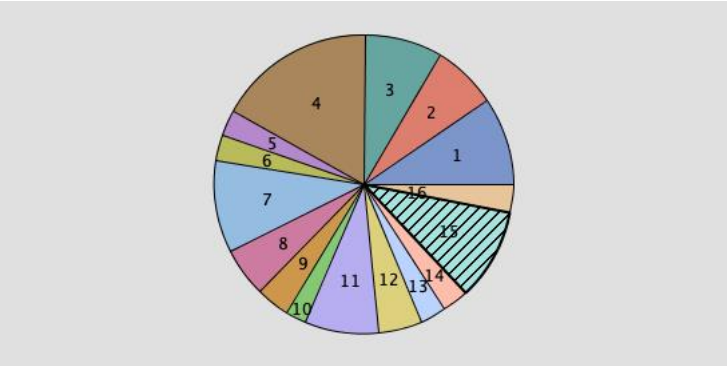


Figure 4: Cluster Frequency Chart

5) **Text filtering** is an information seeking process in which documents are selected from a dynamic **text** stream to satisfy a relatively stable and specific information need. The following concept link is generated for reddit_sentiment dataset.

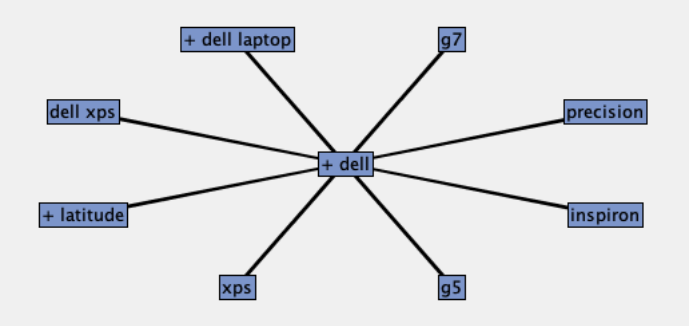


Figure 5: Concept Links

6) The **Segment Profile** node allows to get a better idea of what makes each segment unique or at least different from the population. The node generates various reports that aid in exploring and comparing the distribution of the factors within the segments of dataset samples.

7) Machine learning methods applied :

Decision Tree:

A **Decision Tree** node can be used to classify observations based on the values of nominal, binary, or ordinal targets. It can also predict outcomes for interval targets or the appropriate decision when you specify decision alternatives. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input.

One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it.

The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. In predictive modelling , the decision is the predicted value.

Here we applied two kinds of decision tree types in order to compare. But the result of Training data percentage 20.42% or the rate of validation is almost as same as one another.

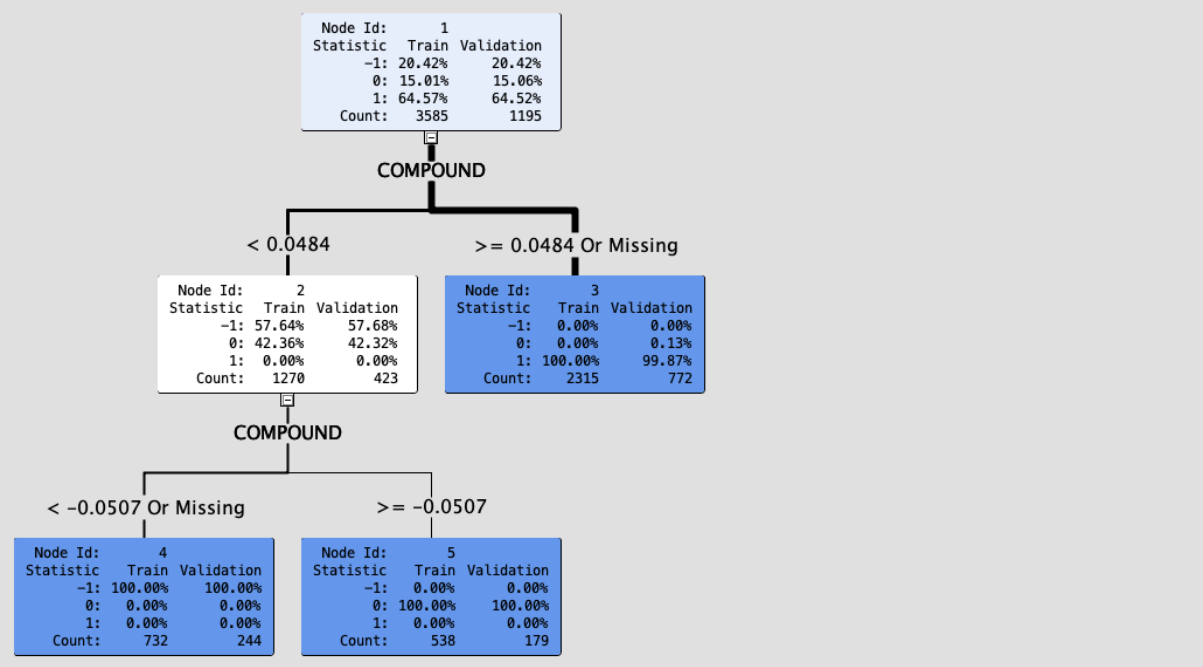


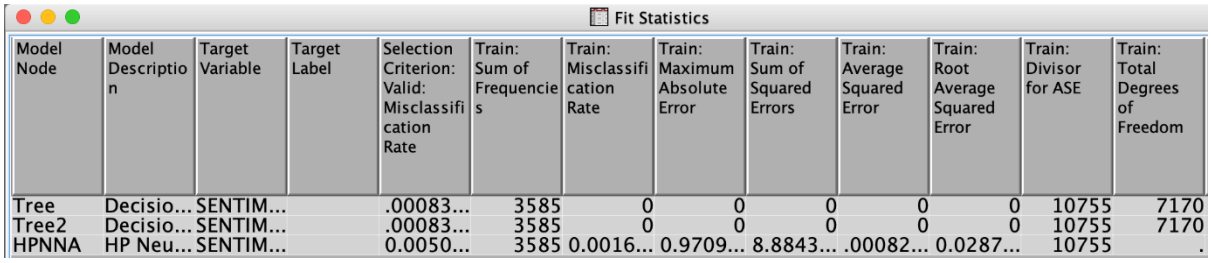
Figure 6: Decision tree

Neural network:

The Hp neural network was also applied on the reddit sentiment data set in order to observe the model comparison and what algorithm helps better to make a better predictive analysis for text data and train well to classify the most talked sentiments people convey every day towards social media for laptop brands.

Model comparison results:

The misclassification rate is lesser in hp neural network with 0.0050 than the decision trees which is 0.0083.



Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom
Tree	Decisio...	SENTIM...		.00083...	3585	0	0	0	0	0	10755	7170
Tree2	Decisio...	SENTIM...		.00083...	3585	0	0	0	0	0	10755	7170
HPNNA	HP Neu...	SENTIM...		0.0050...	3585	0.0016...	0.9709...	8.8843...	.00082...	0.0287...	10755	.

Figure 7: Fit Statistics

The goal is to train the dataset with various algorithm and to observe which model gives better accuracy and classify better, to predict sentiments in general about different brand of laptops on social media.

As we have extracted and constructed dataset for reddit comments and also tweets from twitter. Same procedures as above have been applied for the other two datasets as well.