

# Final Project Report

**Module: NLP**

**Title: Enhancing E-commerce by customer reviews analysis using Text analysis and Natural Language Processing (NLP)**

**Student: Sujatha Ramesh**

**Matri. No: 30006034**

## Executive summary:

### Introduction:

Merchants selling products on the Web often ask their customers to review the products that they have purchased and the associated services. Today, Amazon is becoming more and more popular and the number of customer reviews that a product receives grows rapidly. It is the keys to decision making process to get the customer a better idea about the products. Other than knowing whether the customer is happy or not, the seller can also know how users feel about each feature in the product. Sometimes reviewers write a lot about their lifestyle and the use case they have found for the product as well. Though the “vocal minority” is few, the number of users who are impacted by reviews is large.

A study found that 63% of users prefer online sites that have reviews. Customers who visit review pages have an exciting 105% more chance of buying from the same website (<https://cxl.com/blog/user-generated-reviews/#:~:text=Reevoo%20found%20that%2050%20or,site%20that%20has%20user%20reviews>).

Mining these reviews gives insights to both the online service provider as well as the seller who has listed the product. These insights can later be used in brand communications for the product. It also finds opportunities or gaps in a category and hence get the “voice of the customer” to create a new product or even start a new business.

### Opportunity:

Opinion information is very important for businesses and manufacturers. They often want to know in time what consumers and the public think of their products and services. However, it is not realistic to manually read every post on the website and extract useful viewpoint information from it. There must be too much data. Sentiment analysis allows large-scale processing of data in an efficient and cost-effective manner. In order to know

more about sentiment analysis, this project explores sentiment analysis on business to understand its strengths and limitations.

### **Solution:**

This project used dataset of the **Amazon face mask reviews**, and then built a model to predict the sentiment of the comments by using RapidMiner and machine learning algorithm- Naïve Bayes and decision tree.

### **Methodology:**

Firstly, the sentiment score has been extracted from the customer's review and investigated whether it is positive or negative manually. Then, the confidence level of positivity and negativity has been checked and finally predicted the accuracy using the algorithms.

### **Conclusion:**

As the result, the accuracy of the predicted sentiment in customer's review using decision tree gives a good accuracy compare to Naïve bayes algorithm.

## **PROJECT:**

# **1 Introduction**

This is the **part 1** of the project for the Spring semester module **Text analysis and NLP**. This project covers,

1. Introduction and Motivation.
2. Detailed view of Data including Data pre-processing, Descriptive statistics, Data visualization and Main analysis plan.
3. Extract sentiment.
4. Methodology.
5. Limitations.
6. Conclusion.

## **1.1 Motivation:**

Amazon review analysis is an invaluable advantage that AI and machine learning have given to businesses. Amazon and online shopping are synonymous with each other, more so because the platform has given a chance for companies with modest resources to grow

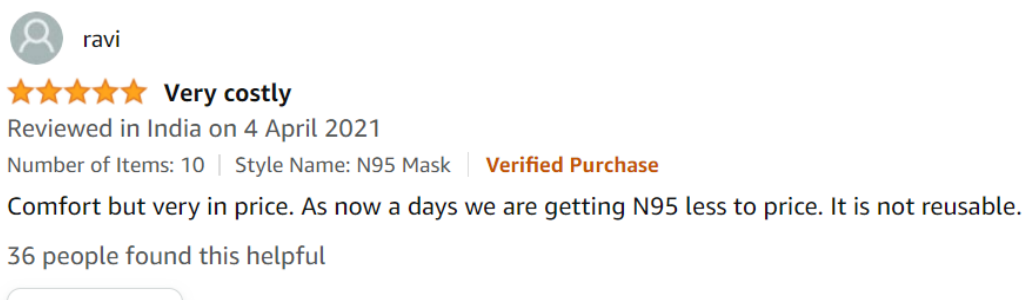
larger than they could. A small company in a remote part of the world can set up an account on Amazon and sell to customers around the globe.

Because of its popularity and ubiquity, Amazon is really the place where people actually spend time and write detailed reviews, unlike other platforms where consumers have to be nudged. Amazon review data analysis can tell companies a lot about their product, even elements that they might not have thought of.

A sentiment analysis in natural language processing (NLP) tasks to not only identify aspects of the products from the Amazon reviews but also enable brands to look beyond star ratings. Amazon review data analysis can give insightful customer information that can be harnessed for product betterment. For this example, I choose to analyze a popular branded face mask that has lot of comments and reviews on Amazon.

Amazon review data analysis is invaluable to brands because the reviews are written by actual users of the product, which means, what's written is usually a first-hand experience. This data is a treasure trove for brands because people can be very vocal about their experience and give key information about the positives and negatives of a product, including its delivery and the customer service they received. This has also unfortunately lead to fake reviews spearheaded by companies. This is an unethical practice which shortchanges the public.

Another important reason is that even though a product may have received 5 stars, it doesn't necessarily mean that the product was good. Digging deeper into the comments, many a time, shows that there are many negative.



In the above case, even though the customer is not 100% satisfied with his experience but, the rating for his comment is given as 5 stars.

## 1.2 Aim:

The aim of this research project is to analyze the sentiment analysis including positivity and negativity of the customer reviews and investigate the accuracy of the sentiment score using naïve bayes and decision tree algorithms.

### **1.3 Research questions:**

1. How much positivity and negativity can be found from each reviews of the customer using sentiment analysis?
2. Is the confidence level of positivity and negativity shows good accuracy in prediction?

## **2 Data:**

I choose Amazon face mask's customer reviews for sentiment analysis whether the customer have a positive or negative impact on the particular product brand.

### **Social and research view why I choose customer reviews on face mask for sentiment analysis:**

Today face mask has become a part of the human life and it is mandatory to wear in public transports and shops since 2020. The goal of most governments is to stop COVID-19 from spreading. Obviously, one particular measure to reduce the spread is to wear face mask. From the evidence of wearing mask, can understand how such a measure contributes in reducing infections. This research article shows how face mask significantly reduced COVID-19 cases in Germany. ( <https://www.pnas.org/doi/10.1073/pnas.2015954117> ).

### **Hypothesis:**

The reviews on product at Amazon depends on the customer satisfaction or their frustration for the particular product. I would like to enhance my study of the customer's opinion on face mask they buy and observe how customer feels about the quality or the nature of the product as it is important to human life today.

### **2.1 DATA RETRIEVAL AND UNDERSTANDING:**

The dataset is extracted from the Amazon product page using the extension called Amazon reviews Exporter.

#### **Dataset size:**

I collected 10,067 reviews totally. After data pre-processing, dataset has 8,869 reviews.

#### **Data Exploration:**

This dataset has 8 columns namely id, profilename, text, date, title, rating, images and helpful.

ExampleSet (Retrieve)

Open in

Turbo Prep

Auto Model

Filter (10,067 / 10,067 examples): all

Row No.	id	profileName	text	date	title	rating	images	helpful
1	R12R75KL44...	Abhishek Raj	I found the p...	Reviewed i...	Good Product	5	https://image...	254
2	R1JBVRPYU...	tejas	Masks are t...	Reviewed i...	Comfortable ...	5	https://image...	205
3	R24A67ANC...	Arpit	This CERTI...	Reviewed i...	awesome N9...	5	https://image...	107
4	R2XEP5P47X...	Rajni gupta	This is a nic...	Reviewed i...	Great mask	5	https://image...	88
5	RE73LQOE8...	Sunil jangir	Ã Mask bah...	Reviewed i...	Acha he or co...	5	https://image...	85
6	R3CT5SP7C...	Avirupa Chak...	This pande...	Reviewed i...	Soft inner lini...	5	https://image...	95
7	R3NTONE3V...	prasenjit sing...	Ã nice	Reviewed i...	nice	5	?	81
8	R1EG74FYN...	Amazon Cust...	Giving it to 1 ...	Reviewed i...	Not recomme...	1	?	77
9	R2RR012SJ...	kapil	Very good q...	Reviewed i...	As expected	5	https://image...	76
10	RTVJHQWFS...	sanket mahale	Soft and hig...	Reviewed i...	India made g...	5	https://image...	76
11	R1DM9TV8L...	Dhaval Trivedi	Mask is goo...	Reviewed i...	Greedy comp...	1	?	39
12	RIX68ZZGU9...	PreethamR	The mask h...	Reviewed i...	Do not buy !	1	https://image...	32

**Id** – It contains the identification number of the customer.

**Profilename** – It is the name of the customer.

**Text** – It contains reviews of the product given by the customer.

**Date** – This column contains the reviewed date of the product.

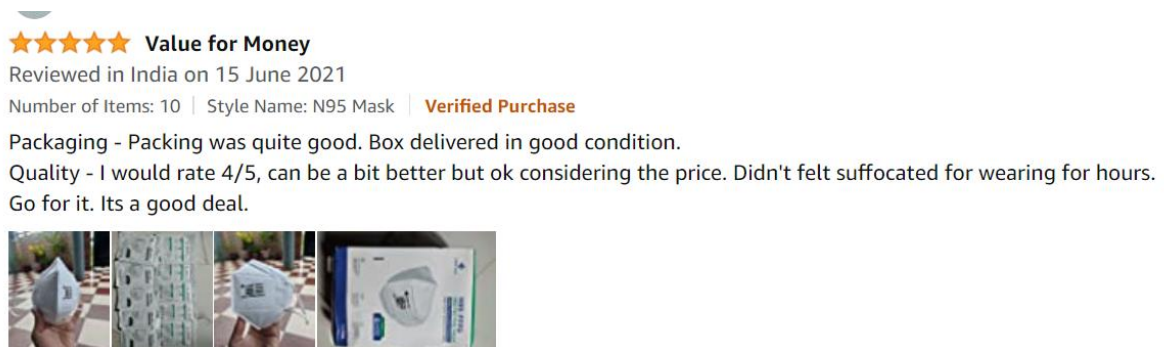
**Title** – This title column contains the short description of the reviews.

**Rating** – It contains the rating for the product ranges from 1 – 5 given by customer.

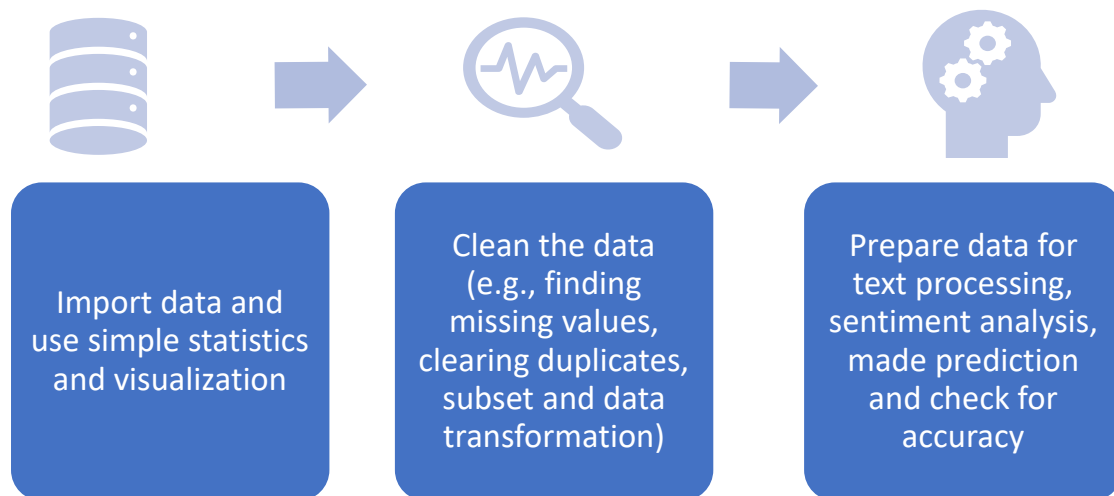
**Images** – This column contains the images of the product which is brought by the customer.

**Helpful** – This column contains the number of the people who find that the reviews are helpful.

In the below picture, can find the example of face mask review containing images, ratings, reviews, date etc.



## 2.2 Analysis Plan:



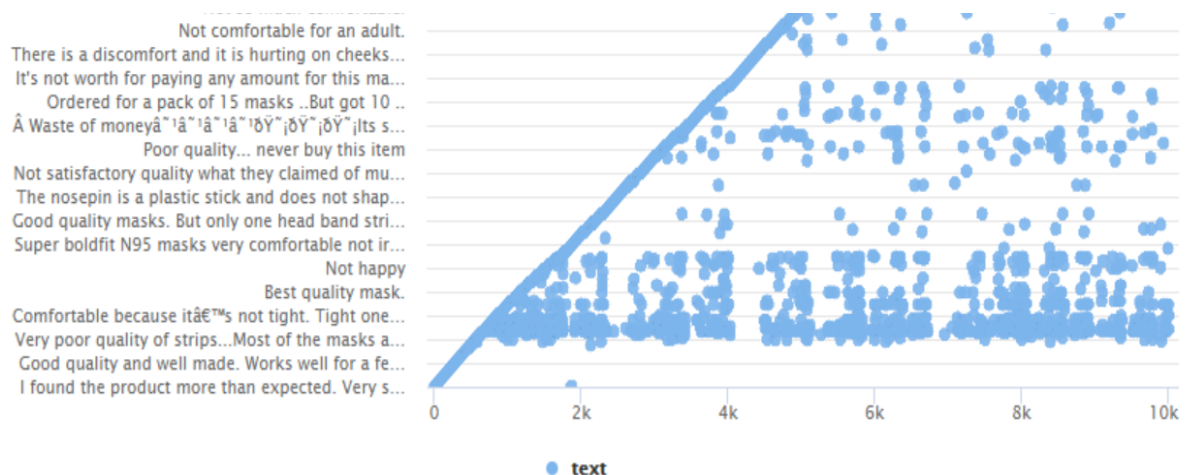
### 2.3 Descriptive statistics:

### 2.3.1 Text attribute:

This is the single variable – quantitative statistics. The attribute “text” has one unknown text review and 207 “Good” reviews. It also has 96 missing values and the type of the text is nominal.

			Least	Most	Values
✓ text	Nominal	96	ðʏ™f (1)	Good (207)	Good (207), Good product (85)

The below visualization shows the distribution of text attribute.

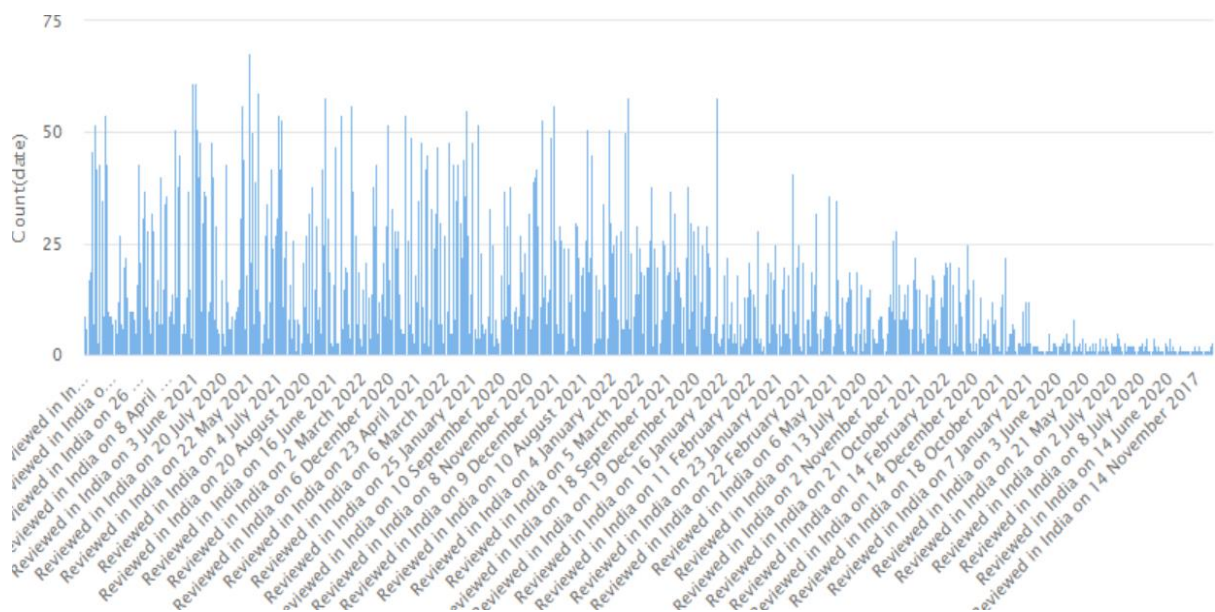


### 2.3.2 Date attribute:

The date attribute has the date of the review given by the customer. From the below statistics it shows that the date of the first comment for this product is given on 14<sup>th</sup> November 2017 and the latest given on 1<sup>st</sup> March 2021 in India.

▼ date	Nominal	0	Least Reviewed [...] 2017 (1)	Most Reviewed [...] 2021 (68)	Values Reviewed [...] June 2021 (68), Re
--------	---------	---	----------------------------------	----------------------------------	---

The Visualization of this statistics can be seen below. It shows that the most reviewed day is 22<sup>nd</sup> May 2021 and less reviewed in mid of 2020s.

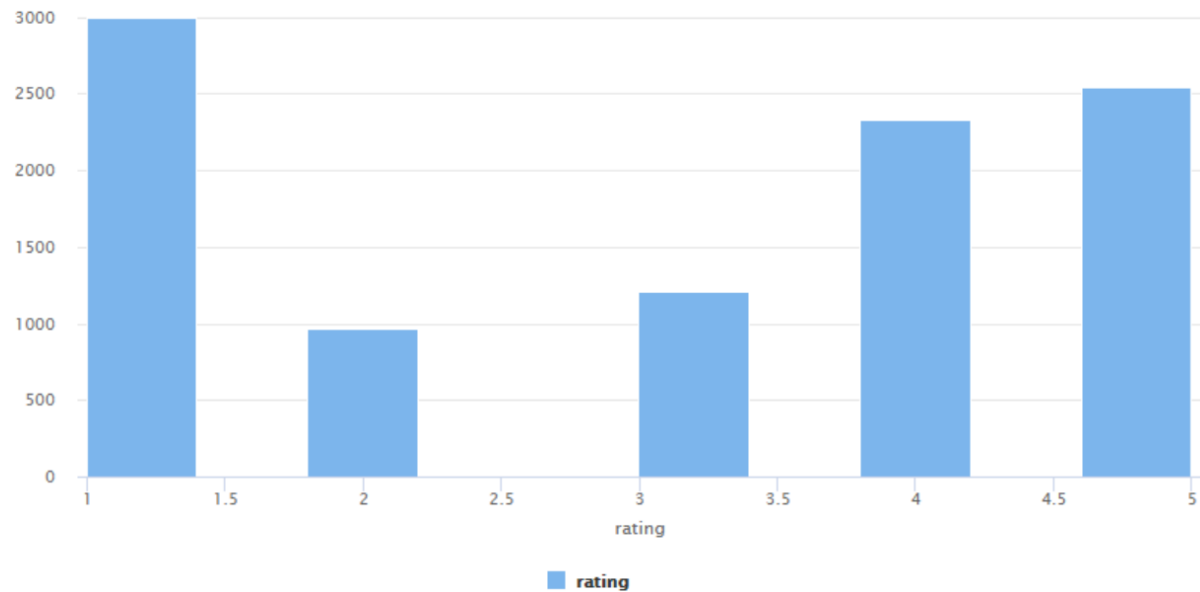


### 2.3.3 Rating:

The rating ranges from 1 to 5. It is categorical and in the statistics shows the minimum rating is 1 and the maximim rating is 5 with the average of 3.044.

▼ rating	Integer	0	Min 1	Max 5	Average 3.044
----------	---------	---	----------	----------	------------------

And the visualization as shown below.

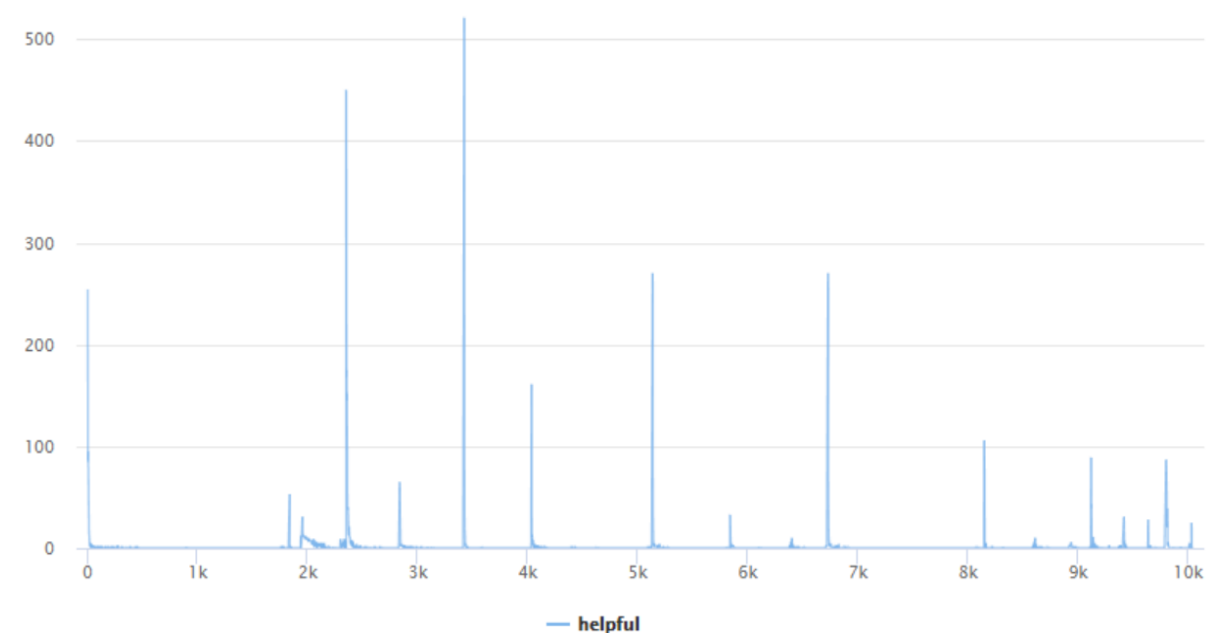


#### 2.3.4 Helpful column attribute:

The statistics shows the minimum of zero customer find the reviews helpful and the maximum of 521 customers find helpful with the average of 1.215.

✓ helpful	Integer	0	Min	Max	Average
			0	521	1.215

From the below line plot, around 3.5k people find the comment most helpful.



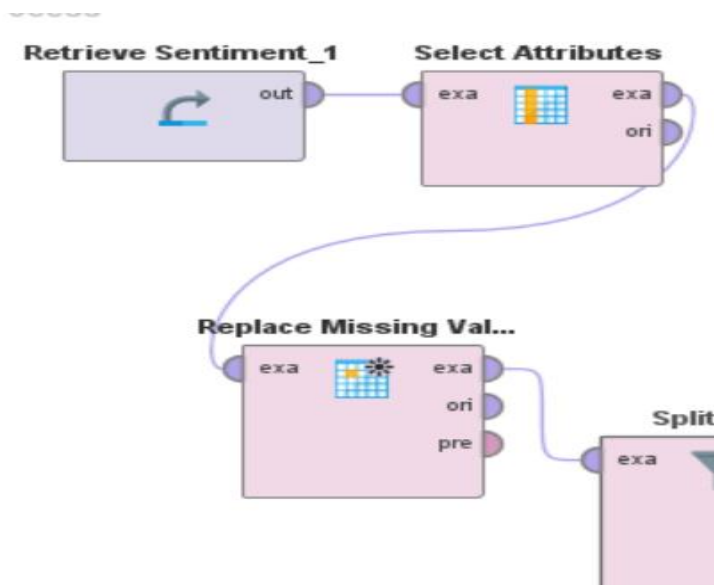


**2.4 Tools used in this project:** Rapidminer, Power BI for Visualization.

## 2.5 Data Pre-processing:

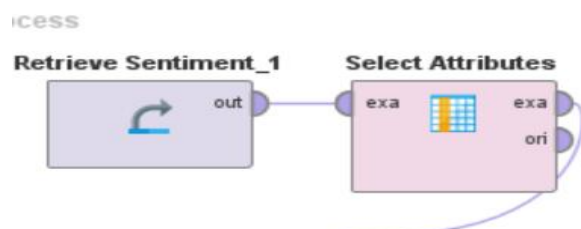
Data pre-processing is the process of transforming raw data into an understandable format prior to processing and analysis. It involves in reformatting data, making corrections to data and the combining of datasets to enrich data.

This careful and comprehensive data preparation ensures analysis to make the data more accurate and meaningful. Firstly, the data can be retrieved by using “retrieve” operator in the RapidMiner.



### 2.5.1 Select Attributes Operator:

This operator helps us to subset the dataset.



### 2.5.2 Missing value treatment:

This dataset contains 96 missing values in text attribute.

ExampleSet (Select Attributes)		ExampleSet (/Local Repository/sentiment_extract_data)				
Name	Type	Missing	Statisti...	Filter (2 / 2 attributes):	Search for Attributes	
✓ text	Polynomial	96	Least Squares (1)	Most Good (207)	Value	Good
✓ rating	Integer	0	Min 1	Max 5	Average	3.0

### 2.5.3 Replace Missing Values:

This operator helps to replace the missing values by minimum, maximum and average value of that Attribute. After filling the missing values,

ExampleSet (Replace Missing Values)		ExampleSet (/Local Repository/sentiment_extract_data)				
Name	Type	Missing	Statisti...	Filter (2 / 2 attributes):	Search for Attributes	
✓ text	Polynomial	0	Least Squares (1)	Most Good (303)	Value	Good
✓ rating	Integer	0	Min 1	Max 5	Average	3.044

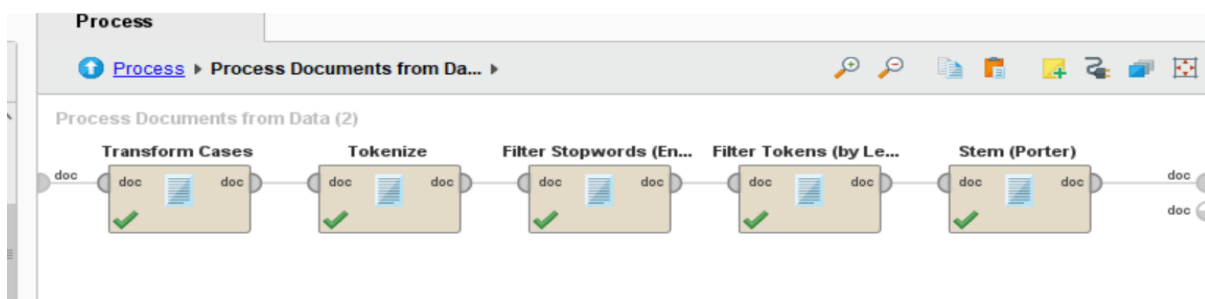
### 2.5.4 Text Processing:

Many operators in RapidMiner can contain nested operators. Double click the “Process Documents from data” operator will bring to a new empty process panel.

#### Process document from data:

RapidMiner has a large set of text processing operators available. Here I use five of them to create a word vector list. Again, use the operator search box to quickly find each operator. The purpose of “Process Document from Data” operator is to extract informations from the structured content of a document.

From the below picture, Inside the operator “Process Documents from Data”, the text processing can be done by performing multiple steps.



### **Transform cases:**

This operator transforms upper case of the text documents to lower case and vice versa.

### **Tokenize:**

Tokenize is an operator for splitting the sentence in the document into a sequence of words.

### **Filter Stopwords(English):**

This operator removes common English words such as 'a' and 'the' etc. Word like these are very noisy unless removed.

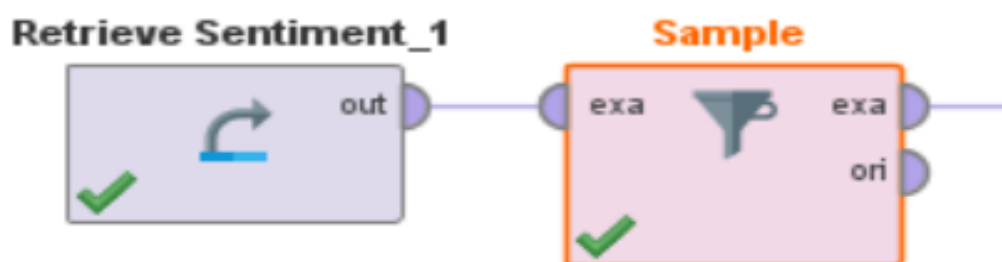
### **Filter Tokens(by Length):**

This operator filters tokens based on their length. i.e the min chars and max chars can be given to filter the tokens by length.

### **Stem (Porter):**

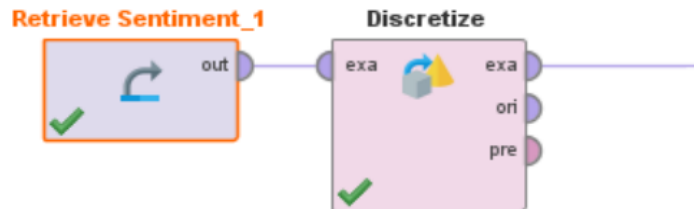
stemming is a very important concept in natural language parsing. It allows one to reduce words to their base or stem. The aim of stemming is to reduce related forms of a word to a common base form. i.e fishing", "fished", "fish", and "fisher" to the base word, "fish". One of the most popular stemmer is the Porter stemmer.

### **2.5.5 Sample Operator:**



This sample operator is used to select the sample from the data by selecting the sample size. For example, if sample size is 10, it gives the result of displaying 10 samples from the dataset.

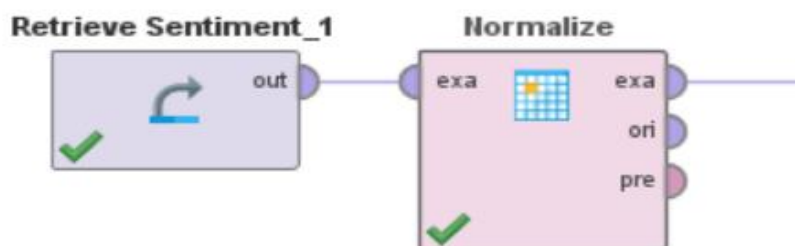
### 2.5.6 Discretize Operator:



This Discretize operator is used to put the values in range. The number of bins is given as 3. So, it gives the nominal value of ranges 3.

Index	Nominal value	Absolute count	Fraction
1	range1 $[-\infty - 9.500]$	3542	0.353
2	range2 $[9.500 - 23.500]$	3276	0.326
3	range3 $[23.500 - \infty]$	3229	0.321

### 2.5.7 Normalize Operator:



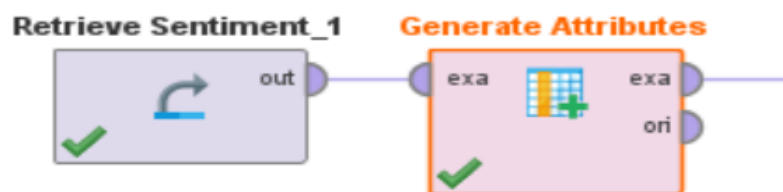
The normalize operator is used to normalize the value. Here, I tried to normalize the value of positivity giving minimum and maximum values of 0 and 1. This is called min-max normalization. In the below table, it shows the value of positivity ranges from 0 and 1.

Open in [Turbo Prep](#) [Auto Model](#)

Row No.	Positivity	id	profileName	text	date	title
1	0.142	R12R75KL44...	Abhishek Raj	I found the p...	Reviewed in I...	Goo
2	0.189	R1JBVRPYU...	tejas	Masks are t...	Reviewed in I...	Con
3	0.557	R24A67ANC...	Arpit	This CERTI...	Reviewed in I...	awe
4	0.383	R2XEP5P47X...	Rajni gupta	This is a nic...	Reviewed in I...	Grea
5	0.041	RE73LQOE8...	Sunil jangir	Â Mask bah...	Reviewed in I...	Acha
6	0.415	R3CT5SP7C...	Avirupa Chak...	This pande...	Reviewed in I...	Soft
7	0	R3NTONE3V...	prasenjit sing...	Â nice	Reviewed in I...	nice
8	0.120	R1EG74FYN...	Amazon Cust...	Giving it to 1 ...	Reviewed in I...	Not
9	0.257	R2RR012SJ...	kapil	Very good q...	Reviewed in I...	As e
10	0.104	RTVJHQWFS...	sanket mahale	Soft and hig...	Reviewed in I...	Indi
11	0.104	R1DM9TV8L...	Dhaval Trivedi	Mask is goo...	Reviewed in I...	Grea
12	0.038	RIX68ZZGU9...	PreethamR	The mask h...	Reviewed in I...	Do r
13	0.178	R3P1VD0Z1...	Jayanth	Don't go by t...	Reviewed in I...	Don

## 2.5.8 Generate Attributes Operator:

In this data transformation method, a new column can be generated using “Generate Attributes” operator.



Here, in the function descriptions the expressions should be given. For example, I wanted to create a column named “rating performance” says the rating 5 as Excellent, 4 as Good, 3 as Not bad and so on. The expression used is - `if(rating == "1","Very bad", if(rating == "2","Bad",if(rating == "3","Not bad", if(rating == "4","Good",if(rating == "5","Excellent","")))))`

From the below picture, it is showed that the new column is created with comments given to the ratings.

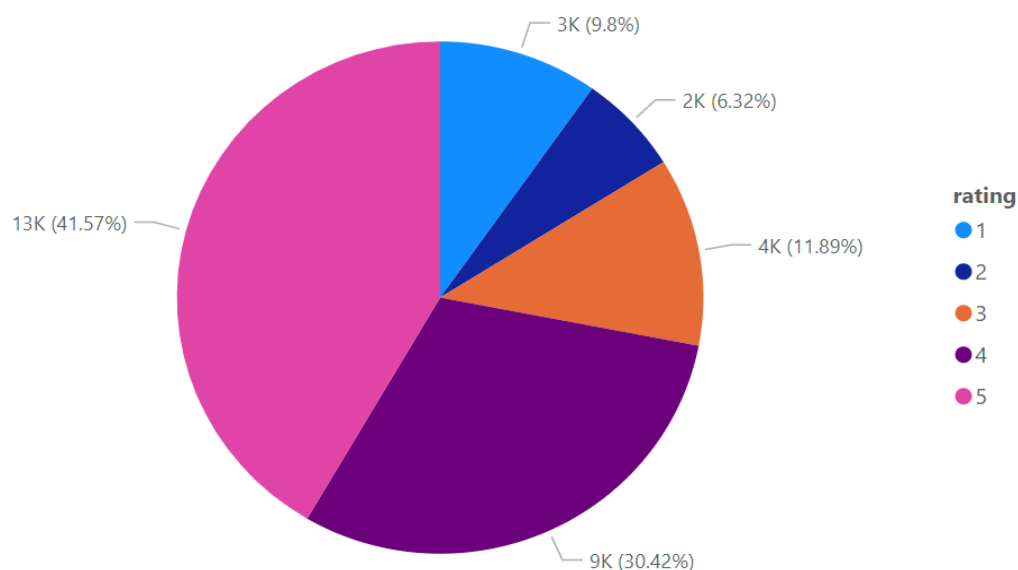
ExampleSet (Generate Attributes)				
ExampleSet (//Local Repository/sentiment_extract_data)				
Open in		Turbo Prep	Auto Model	
Filter (8,869 / 8,869 examples): all				
Row No.	text	id	rating	rating perfor...
1	I found the pr...	R12R75KL44...	5	Excellent
2	Masks are to...	R1JBVRPYU...	5	Excellent
3	This CERTIFI...	R24A67ANC...	5	Excellent
4	This is a nice...	R2XEP5P47X...	5	Excellent
5	Â Mask bahut...	RE73LQOE8...	5	Excellent
6	This pandem...	R3CT5SP7C...	5	Excellent
7	Â nice	R3NTONE3V...	5	Excellent
8	Giving it to 1 ...	R1EG74FYN...	1	Very bad
9	Very good qu...	R2RR012SJ...	5	Excellent
10	Soft and high ...	RTVJHQWFS...	5	Excellent
11	Mask is good...	R1DM9TV8L...	1	Very bad
12	The mask ha...	RIX68ZZGU9...	1	Very bad
13	Don't go by th...	R3P1VD0Z1...	1	Very bad

## 2.6 DATA VISUALIZATION:

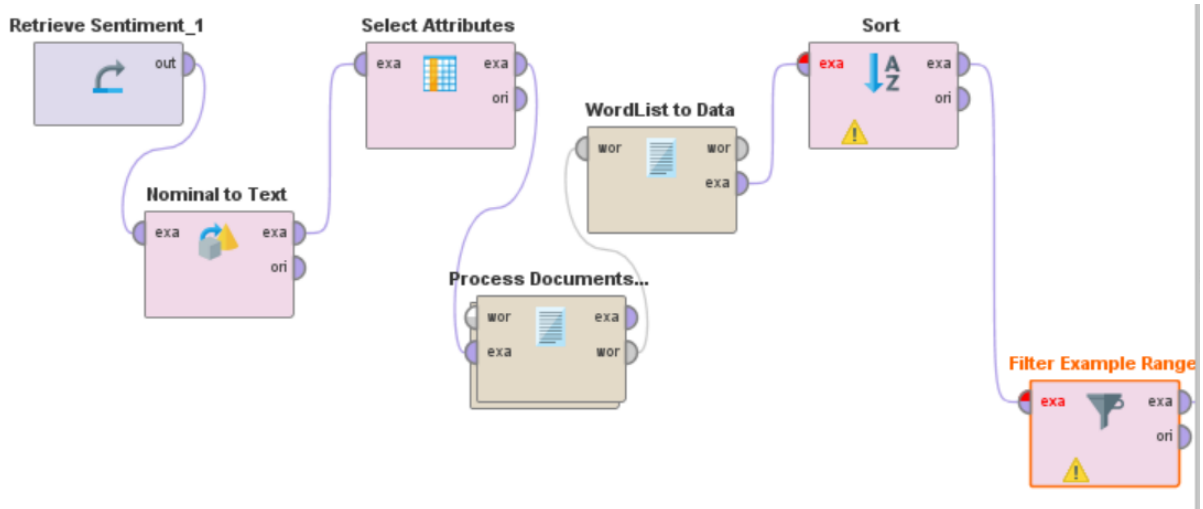
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

### 2.6.1 Pie chart :

Here, I would like to visualize the customer's rating using **PowerBI**. It also shows the total percentage of the ratings individually from 1 to 5.



### 2.6.2 Word cloud:



**WordList to data-** operator helps to get the total occurrences of the word list from the text attribute. Here, it is the column “total”.

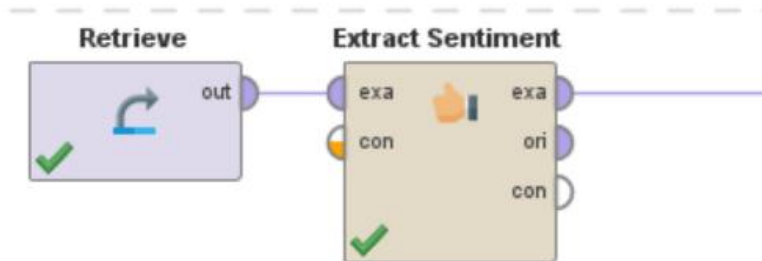
**Sort operator-** sort-by the words from A-Z with the attribute to descending order. Here, I sorted the column “total” to descending order to get the top occurrences.

**Filter Example Range** operator filters the total data range from first example to last example. Here, I gave 1 to 50. So, it displayed about 50 words in the wordcloud plot type.

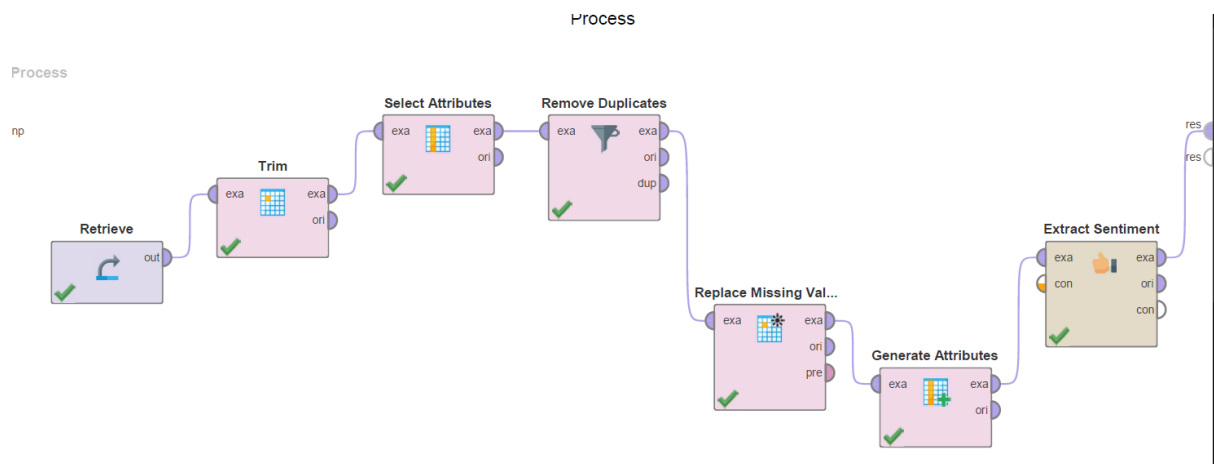


### 3 Extract Sentiment:

This operator is used in text processing. Here, I used the model “Vader” and in the text attribute “text” to extract the sentiment from it. The raw dataset is used for this.



The process includes pre-processing with the extract sentiment below,



The **Trim operator** used below is used to delete any blanks in the document. After pre-processing the dataset, Extract sentiment is applied to it.

After applying “extract sentiment” to the process, it extracts the value of score, Scoring string, negativity, positivity, uncovered tokens and Total tokens from the reviews of the customer.



ExampleSet (Generate Attributes)		ExampleSet (/Local Repository/sentiment_extract_data)						
Open in		Turbo Prep	Auto Model	Filter (8,869 / 8,869 examples):				
Row No.	Score	Scoring Stri...	Negativity	Positivity	Uncovered T...	Total Tokens	text	id
1	-0.590	nice (0.46) e...	1.923	1.333	38	46	I found the pr...	R12R75KJ
2	1.462	comfortable (...)	0.308	1.769	30	34	Masks are to...	R1JBVRP
3	5.231	competitive (...)	0	5.231	84	96	This CERTIFI...	R24A67AN
4	3.590	nice (0.46) c...	0	3.590	113	119	This is a nice...	R2XEP5P
5	0.385	comfort (0.38)	0	0.385	27	28	Â Mask bahut...	RE73LQO
6	3.897	compelled (0....)	0	3.897	130	139	This pandem...	R3CT5SP
7	0		0	0	1	1	Â nice	R3NTONE
8	0.564	giving (0.36) ...	0.564	1.128	52	57	Giving it to 1 ...	R1EG74F
9	2.410	good (0.49) ...	0	2.410	34	40	Very good qu...	R2RR012
10	0.974	good (0.49) ...	0	0.974	50	52	Soft and high ...	RTVJHQW
11	0.051	good (0.49) ...	0.923	0.974	47	51	Mask is good...	R1DM9TV
12	-0.282	easily (0.36) ...	0.641	0.359	42	44	The mask ha...	RIX68ZZG
13	0.487	forget (-0.23) ...	1.179	1.667	84	92	Don't ao by th...	R3P1VD0

From the above figure,

**Score** – This is the column shows whether the text reviews are positive or negative. If the review is positive, the score is positive. If the review is negative, it score is negative.

**Scoring String**- This column contains the number of words with its score.

**Negativity**- This column contains the total negative score of the reviews.

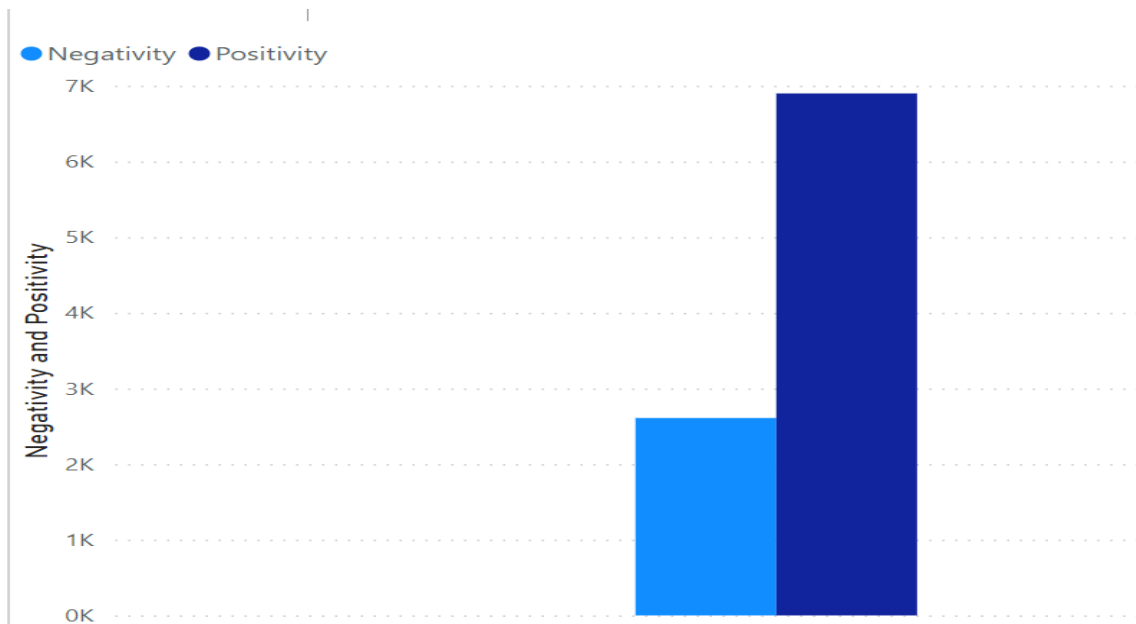
**Positivity**- This column contains the total positive score of the reviews.

**Uncovered Tokens**- This column contains neutral score.

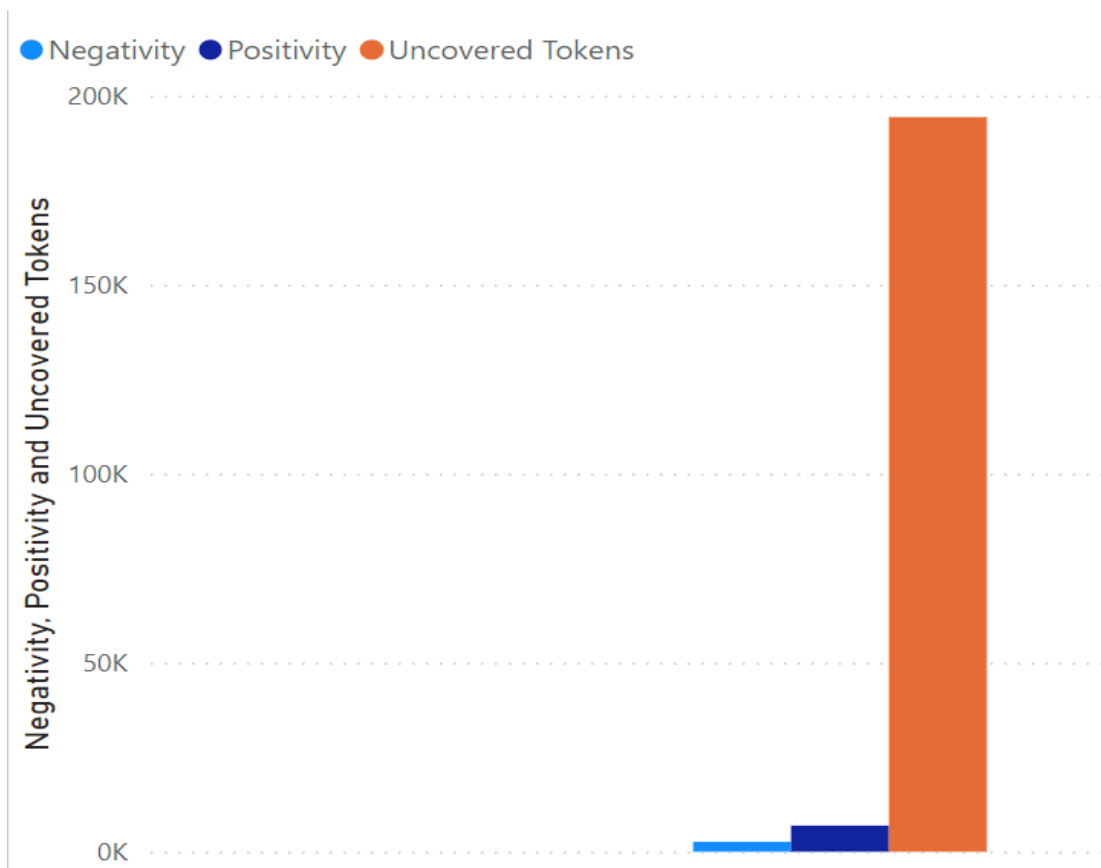
**Total Tokens**- This column contains total number of the tokens that the text or reviews have.

I would like to perform a Data Visualization using PowerBI for this “extract Sentiment” data. The below visualization shows the trend of positivity and negativity of this dataset.

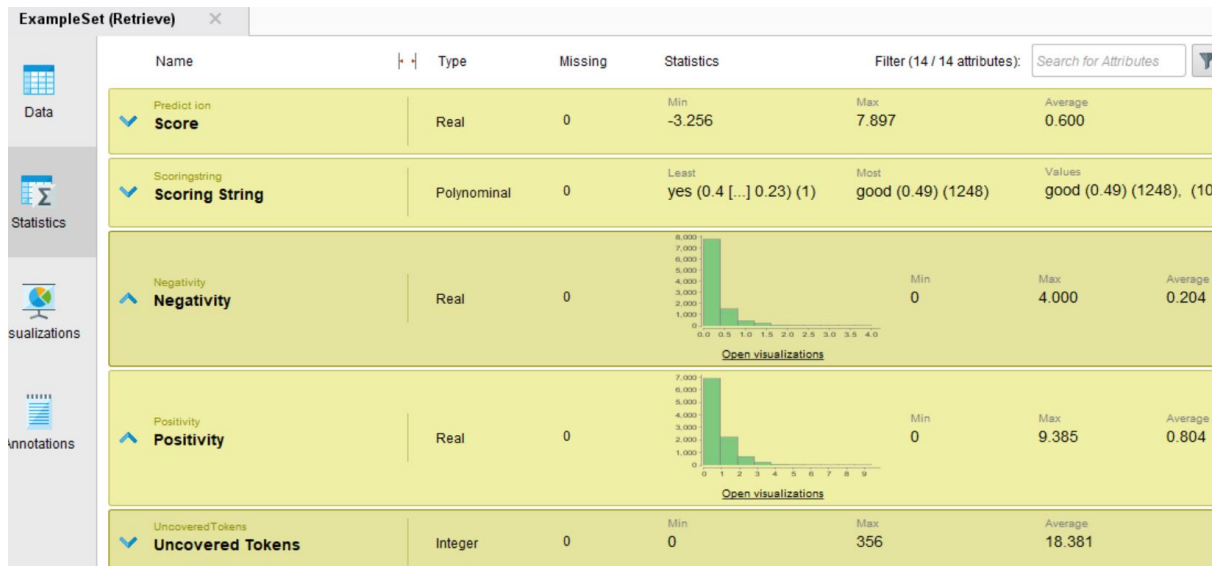
Here, one can understood the positivity in the customer’s reviews is higher that negativity for the product.



Also, I would like to visualise all the three values positivity, negativity and uncovered tokens (neutral). From the below figure, it is shocking that neutral comments are high compared to positive and negative comments. This shows that almost the sentiment analysis trends to positive side.



## 2.7.1 Statistical analysis of extract sentiment data:



The statistical view of the data can provide us a more insights about the dataset. From the above statistics, the score ranges from minimum -3.25 to maximum 7.89. The positivity and negativity show visualize the trend from 0 to 4 and 0 to 9 and uncovered tokens with the maximum value of 356. This is greater than positivity and negativity as I already showed in the above visualization.

## 4 Methodology:

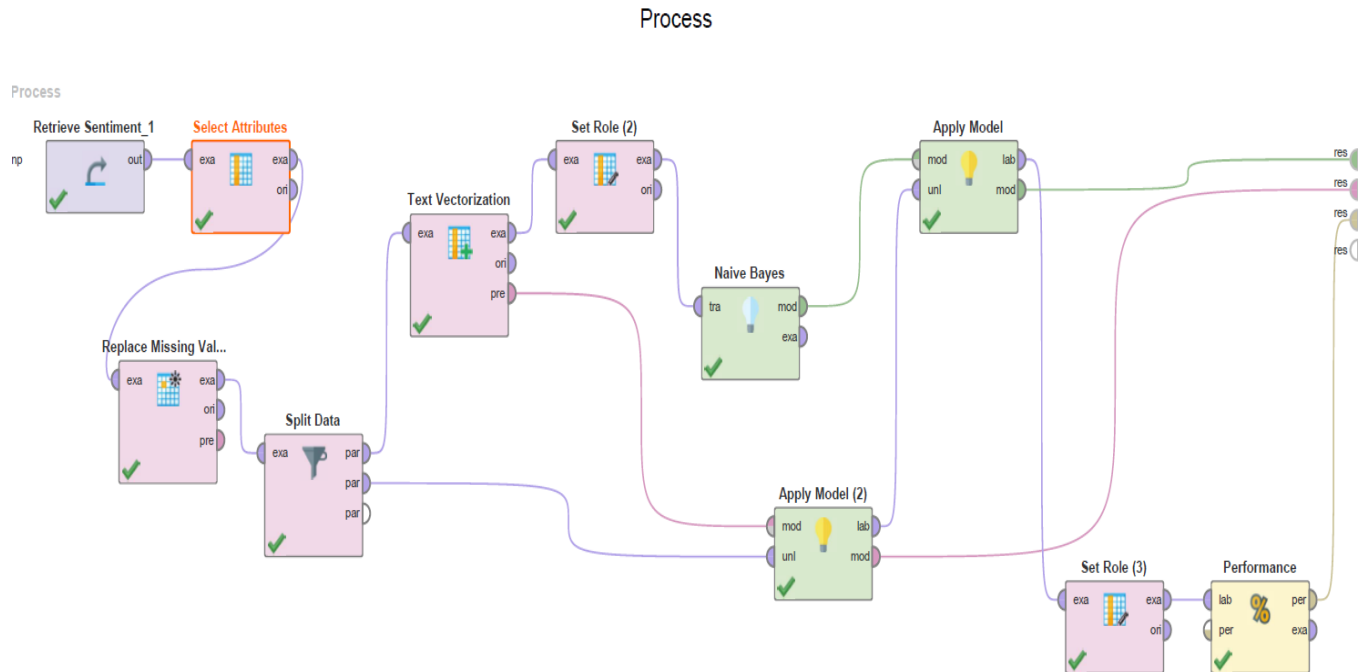
As discussed before, the naïve bayes and decision tree algorithm is used for prediction and perform accuracy.

### 4.1 Naïve bayes algorithm:

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. The pros and cons of naïve bayes algorithm is that the assumption that all features are independent makes naive bayes algorithm very fast compared to complicated algorithms. In some cases, speed is preferred over higher accuracy. It works well with high-dimensional data i.e., text classification. The data used here is extract sentiment data.

## Process:

The whole process to analyse the prediction and accuracy for Naïve bayes



## Process Explanation:

Firstly, I extracted sentiment from the dataset which I explained in “Extract sentiment” (topic 2.7) previously. It results in producing new columns namely score, scoring string, positivity, negativity, uncovered token and total tokens with the existing columns in the dataset. In this extract sentiment dataset i.e. (named as Retrieve Sentiment\_1 in the process), the “Score” column shows whether the text reviews are positive or negative. If the review is positive, the score is in positive number. If the review is negative, the score is in negative number. Here, I added the column named “Sentiment” (6<sup>th</sup> column shows in the below excel figure), this column is added manually and gave a value as “neg” for negative scores and zeros, “pos” for all other positive and neutral scores.

A	B	C	D	E	F	G	H	I	J	K
d	profileName	text	date	title	Sentiment	rating	helpful	Score	Scoring String	Negati
12R75KL	Abhishek	I found the product	Reviewed	Good Product	neg	5.0	254.0	-0.6	nice (0.46) easy (0.	
1JBVRPYI	tejas	Masks are too com	Reviewed	Comfortable and h	pos	5.0	205.0	1.5	comfortable (0.59)	
24A67AN	Arpit	This CERTIFIED N9	Reviewed	awesome N95 mas	pos	5.0	107.0	5.2	competitive (0.18)	
2XEP5P4	Rajni gupt	This is a nice mask	Reviewed	Great mask	pos	5.0	88.0	3.6	nice (0.46) comfor	
E73LQOE	Sunil jangi	Â Mask bahut hi c	Reviewed	Acha he or comfort	pos	5.0	85.0	0.4	comfort (0.38)	
3CT5SP7	Avirupa Cl	This pandemic situ	Reviewed	Soft inner lining an	pos	5.0	95.0	3.9	compelled (0.05) t	
3NTONE	prasenjit s	Â nice	Reviewed	nice	neg	5.0	81.0	0.0		
1EG74FY	Amazon C	Giving it to 1 star n	Reviewed	Not recommended	pos	1.0	77.0	0.6	giving (0.36) dignit	
2RR012S	kapil	Very good quality i	Reviewed	As expected	pos	5.0	76.0	2.4	good (0.49) good (	
TVJHQW	sanket ma	Soft and high qual	Reviewed	India made good p	pos	5.0	76.0	1.0	good (0.49) good (	
1DM9TV	Dhaval Tri	Mask is good but c	Reviewed	Greedy company b	pos	1.0	39.0	0.1	good (0.49) unethi	
1X68ZZGL	Preetham	The mask has big g	Reviewed	Do not buy !	neg	1.0	32.0	-0.3	easily (0.36) bad (-	
3P1VD0Z	Jayanth	Don't go by the sel	Reviewed	Don't buy this. The	pos	1.0	30.0	0.5	forget (-0.23) won	
9BFKUSG	mahadev	I had got same ma	Reviewed	Terrible business st	neg	3.0	15.0	-1.1	amazon (0.18) goc	
1576GES	Pradeep N	Â Good	Reviewed	GoodðŸŽ	neg	4.0	13.0	0.0		
3CMMQFA						4.0	13.0	0.0		

This helps me in finding the accuracy how much the score of the review is accurate. And how much the positivity and negativity of the customer's review on face mask is accurate.

## Operators used in the above process:

**Split Data Operator-** This Operator splits the data into train and test. Here, I split the data into 80% and 20%.

**Text Vectorization Operator-** It is the process of converting text into numerical representation. It is also called as word embeddings. Here, in this project, the text vectorization is applied to the attribute "text" with the document class attribute "Sentiment".

## Outcome:

From the split data the train set connected to Naïve bayes and test set to the model, the result we get is amazing. It returned prediction value for the Sentiment column and the confidence value for positive and negative.

Open in [Turbo Prep](#) [Auto Model](#) File (2,009 / 2,009 examples) [all](#)

Row No.	prediction(Sentiment)	confidence(neg)	confidence(pos)	id	Sentiment	Score	text:table	text:about
1	pos	0	1	R1JBVRPYU...	pos	1.462	0	0
2	neg	0.981	0.019	R9BFKUSGA...	neg	-1.103	0	0
3	neg	0.981	0.019	R16MMQD2L...	neg	-1.821	0	0
4	pos	0	1	R2GAZ1ZUD...	pos	0.487	0	0
5	pos	0	1	R1BYDG3VL...	pos	3.846	0	0
6	pos	0	1	R3V9W6YAC...	pos	0.974	0	0
7	neg	0.981	0.019	R2BU10EDH...	neg	-2.436	0	0
8	neg	0.981	0.019	R1EYNAE1A...	neg	-0.590	0	0
9	neg	0.981	0.019	RF98VS1UB...	pos	0	0	0
10	pos	0	1	R3MCM7QFM...	pos	1.564	0	0
11	pos	0	1	R1VCKROFC...	pos	1.256	0	0
12	pos	0	1	R5V8UFPR8...	pos	2.667	0	0

**Prediction (Sentiment):** This is the prediction made by model from the Sentiment column. If the confidence(neg) is high, it gives negative. If the confidence(pos) is high, it gives positive in the prediction.

**Confidence(neg):** This is the confidence level of the model on the negative reviews.

**Confidence(pos):** This is the confidence level of the model on the positive reviews.

## Result:

Finally, the prediction(Sentiment) and Sentiment are analysed to see the accuracy resulted from the performance. As the result the accuracy is 72.13% using Naïve bayes model.

Result History

**PerformanceVector (Performance)** × **TextVectorizationModel (Text Vectorization)** ×

☒ Table View ☐ Plot View

Criterion: accuracy

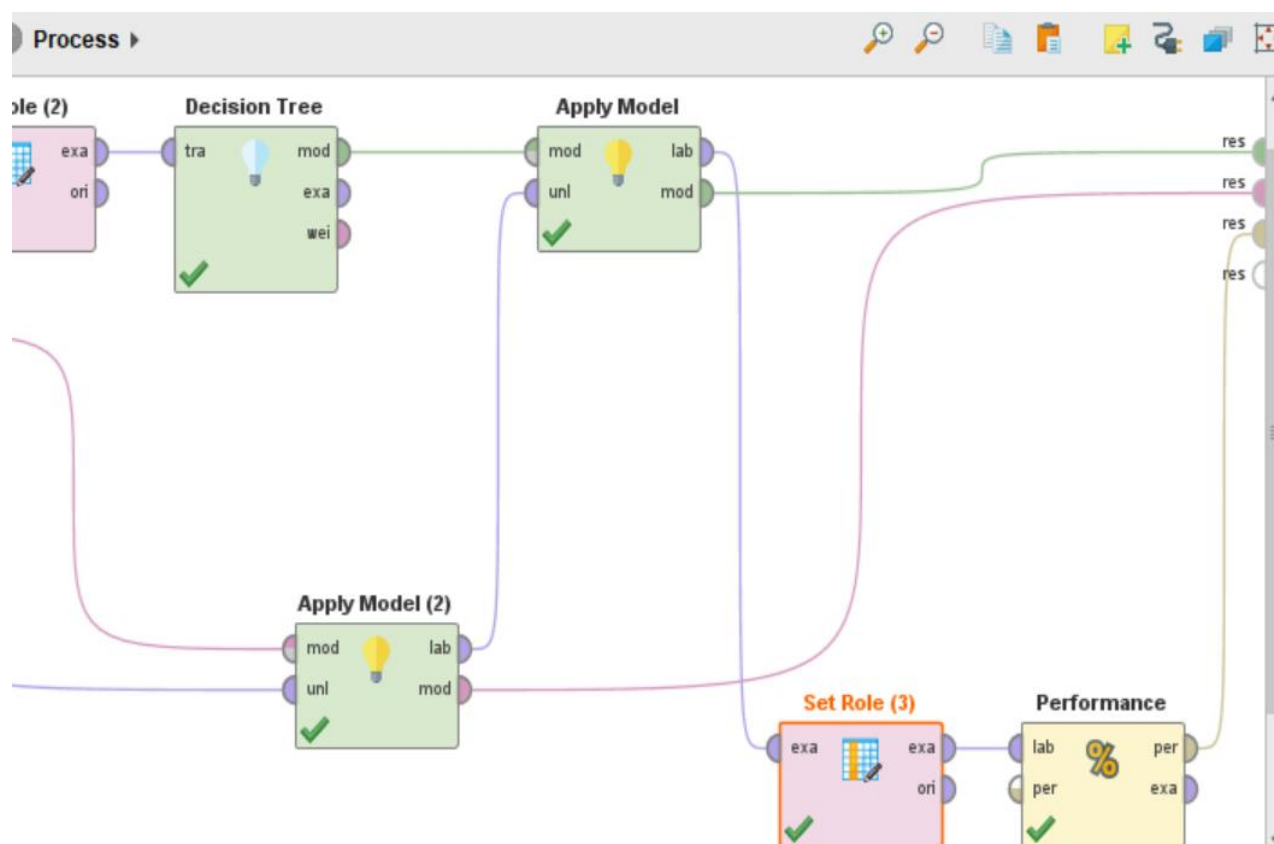
accuracy: 72.13%

	true neg	true pos	class precision
pred_neg	401	240	62.56%
pred_pos	320	1048	76.61%
class recall	55.62%	81.37%	

## 4.2 Decision Tree:

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, it is the purity of the node increases with respect to the target variable.

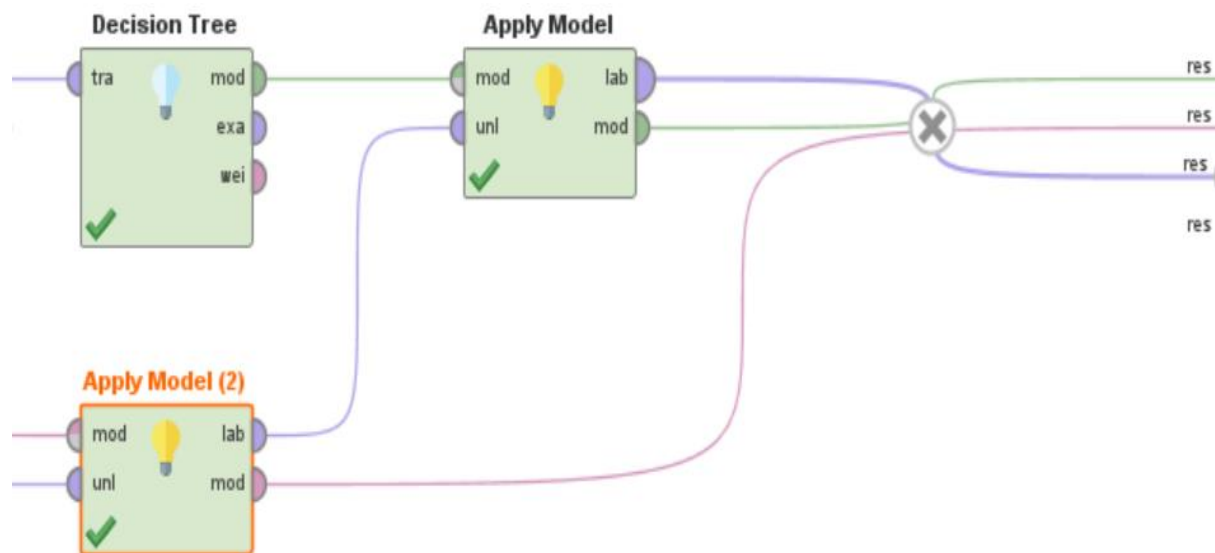
### Process:



### Process Explanation:

As mentioned already (in the process explanation of Naïve bayes) the column "Sentiment" is given manually in the dataset and it is used for prediction of Sentiment column, confidence(neg) and confidence(pos) and finally accuracy is calculated using decision tree algorithm.

## Outcome:



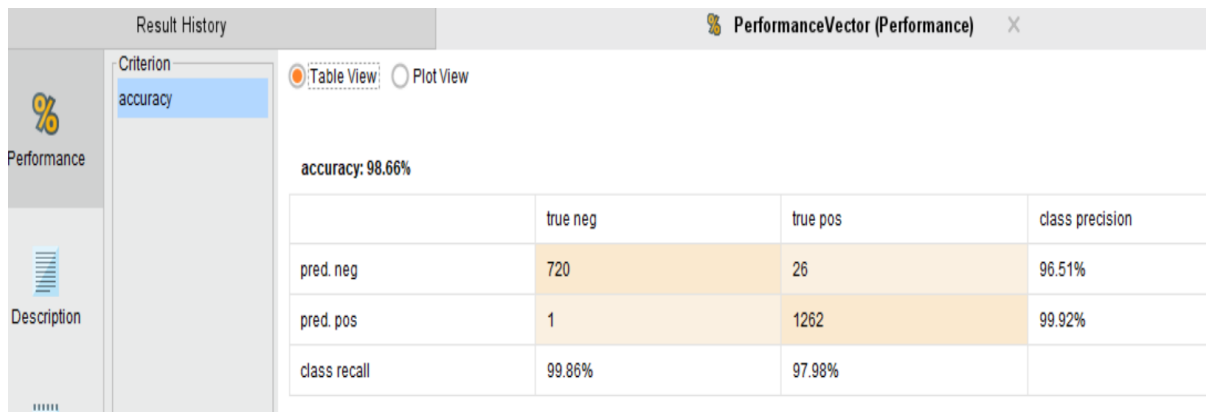
From the split data the train set is connected to Decision tree model and test set is connected to the Apply model, the result we get is, in the below table, it returned prediction value for the “Sentiment” column and the confidence value of positive and negative.

Row No.	prediction(S...	confidence(...	confidence(...	id	Sentiment	Score	textable	textabout
1	pos	0	1	R1JBVRPYU...	pos	1.462	0	0
2	neg	1	0	R9BFKUSGA...	neg	-1.103	0	0
3	pos	0	1	R16MMQD2L...	neg	-1.821	0	0
4	pos	0	1	R2GAZ1ZUD...	pos	0.487	0	0
5	pos	0	1	R1BYDG3VL...	pos	3.846	0	0
6	pos	0	1	R3V9W6YAC...	pos	0.974	0	0
7	neg	1	0	R2BU10EDH...	neg	-2.436	0	0
8	neg	1	0	R1EYNAE1A...	neg	-0.590	0	0
9	pos	0	1	RF98VS1UB...	pos	0	0	0
10	pos	0	1	R3MCM7QFM...	pos	1.564	0	0
11	neg	1	0	R1VCKROFC...	pos	1.256	0	0
12	pos	0	1	R5V8UFPR8...	pos	2.667	0	0
13	neg	0	1	R2VQPLG3D...	neg	0.600	0	0



## Result:

Finally, the prediction(Sentiment) column and Sentiment are analysed to see the accuracy resulted from the performance. As the result, the accuracy of 98.66% performed by decision tree model.



	true neg	true pos	class precision
pred. neg	720	26	96.51%
pred. pos	1	1262	99.92%
class recall	99.86%	97.98%	

## 5 Limitations:

- Only two algorithms are applied in this project: Naïve bayes and Decision tree.
- This project focuses only on prediction and accuracy for the positive and negative sentiment score of customer's reviews.
- Only the reviews of the face mask are used.

## 6 Conclusion:

The aim of this project is now built and is completed successfully with the descriptive statistics, main analysis and methodology as planned. As the result, the Decision tree performed good with the accuracy of 98.66% compared to Naïve bayes. For the future work, the analysis can be done further to get 100% accuracy by using other predictive algorithms like Fast last margin, Generalized Linear model (GLM), SVM, KNN etc

## References

[https://www.researchgate.net/publication/306285022\\_Summarization\\_of\\_Customer\\_Reviews\\_for\\_a\\_Product\\_on\\_a\\_website\\_using\\_Natural\\_Language\\_Processing](https://www.researchgate.net/publication/306285022_Summarization_of_Customer_Reviews_for_a_Product_on_a_website_using_Natural_Language_Processing)

[https://www.researchgate.net/publication/313248238\\_Summarizing\\_customer\\_review\\_based\\_on\\_product\\_feature\\_and\\_opinion](https://www.researchgate.net/publication/313248238_Summarizing_customer_review_based_on_product_feature_and_opinion)

<https://www.analyticsvidhya.com/blog/2021/06/part-5-step-by-step-guide-to-master-nlp-text-vectorization-approaches/>

<https://docs.rapidminer.com/latest/studio/operators/blending/attributes/generation/text-vectorization.html>

<https://aclanthology.org/2020.coling-main.15/>

<https://content.bridgpointeducation.com/curriculum/file/a0a05eaf-474d-49a1-a4c2-ca9a9191f11c/1/Sample%20Executive%20Summary.pdf>

<https://medium.com/analytics-vidhya/decision-tree-with-amazon-food-reviews-5639a7b70cef>

[https://home.uncg.edu/cmp/faculty/j\\_deng/papers/sentiment\\_globecom19.pdf](https://home.uncg.edu/cmp/faculty/j_deng/papers/sentiment_globecom19.pdf)