



---

# مشروع التخرج

---

Walmart-Text Mining/Sentiment Analysis

مقدم إلى مشرف المشروع: د. محمد دويكات، د. نجوان الدلق

فريق التطوير:

بيان بني فضل      ياسمين سلمان      حنين شهاب

2018-2017

جامعة النجاح الوطنية

كلية تكنولوجيا المعلومات – قسم أنظمة المعلومات الإدارية

## Table of Contents

---

<b>1.0 Introduction.....</b>	<b>Error! Bookmark not defined.</b>
1.1Data mining.....	Error! Bookmark not defined.
1.2Text Mining vs. Data Mining:.....	Error! Bookmark not defined.
1.3 Why is Sentiment Analysis important when Analyzing Social Media?Error!	Bookmark not defined.
1.4 flowchart for text mining process .....	Error! Bookmark not defined.
<b>2.0 Tools used in our project ....</b>	<b>Error! Bookmark not defined.</b>
2.1 Rapidminer tool .....	Error! Bookmark not defined.
2.2Facepager tool.....	Error! Bookmark not defined.
<b>3.0 ETL (Extract, Transform, load)for facebook data</b>	<b>Error!</b>
<b>Bookmark not defined.</b>	
<b>4.0 Cleaning Data with excel ....</b>	<b>Error! Bookmark not defined.</b>
<b>5.0 training data process.....</b>	<b>Error! Bookmark not defined.</b>
<b>6.0 Load data by Rapidminer ....</b>	<b>Error! Bookmark not defined.</b>
<b>7.0 ETL (Extract, Transform, load)for twitter data .....</b>	<b><a href="#">50</a></b>
<b>8.0 problems we encountered in the project</b>	<b>Error! Bookmark not</b>
<b>defined.</b>	
<b>9.0 Solutions and recommendations</b>	<b>Error! Bookmark not defined.</b>

# Introduction

---

**Data mining** is the computational process of discovering patterns in large data sets and establish relationships to solve problems through data analysis involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of data mining is to extract information from a data set and transform it into an understandable structure for further use.

## **Text Mining vs. Data Mining:**

**Text mining and data mining** are often used interchangeably to describe how information or data is processed.

data mining, which we can define as the discovery of knowledge from structured data (data contained in structured databases or data warehouses.) Today the majority of available business data is unstructured information (articles, website text, blog posts, etc.). The presence of unstructured information makes it more difficult to effectively perform knowledge management activities using traditional business intelligence tools.

The discovery of knowledge sources that contain text or unstructured information is called “text mining”. So, the main difference between data mining and text mining is that in text mining data is unstructured

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature level, i.e. whether the expressed opinion in a document, a sentence or an entity feature is positive, negative, or neutral.

The accuracy of a sentiment analysis system is, in principle, how well it agrees with human judgments. This is usually measured by precision and recall.

If a program were "right" 100% of the time, humans would still disagree with it about 20% of the time, since they disagree that much about any answer

Our **graduation project is about sentiment analysis**; our idea is to keep handle on how everyone feels about any brand.

For large companies with thousands of daily mentions and comments on social media, news, sites and blogs, it's extremely difficult to do this manually .to combat this problem, sentimental analysis software are necessary. this software's can be used to evaluate the people's sentiment about any brand or personality.

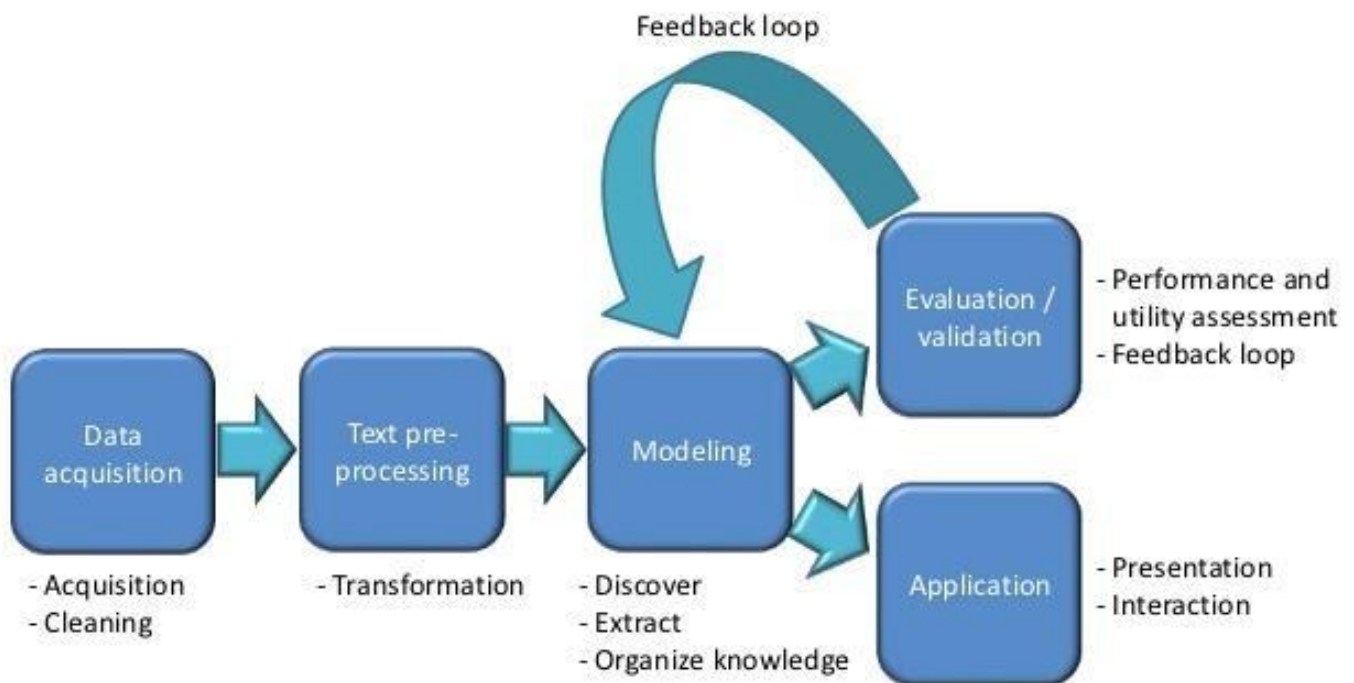
And we choose **Walmart** as a domain to apply sentimental analysis.

**So our problem is** organization development, we want to make sentiment analysis in order to keep up with people opinions, know what they want what they said about our services and products in order to help us to improve our companies, know people trend, control our website and social media.

## Why is Sentiment Analysis important when Analyzing Social Media?

- Determine marketing strategy
- Improve campaign success
- Improve product messaging
- Improve customer service
- Test business KPIs
- Generate leads

## Typical text mining process



## Tools that we used in our project:

---



### **Rapidminer tool:**

is a data science software platform developed by the company of the same name that provides an integrated environment for **data preparation, machine learning, deep learning, text mining, and predictive analytics.**

It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization. RapidMiner is developed on an open core model.

2.



### Facepager tool:

- a tool to simplify the process of gathering data from JSON-based APIs without the use of programming languages or predefined scripts, while leaving large degrees-of-freedom to the user. Thus, we do not restrict any API-endpoints and allow "useless" requests.
- a tool to support the step of "data collection" on a low level
- a tool to document the process of data collection, i.e. errors occurring in the process (on both the side of the API and locally, f.e. ill-defined requests).
- a tool that targets researchers/scientific purposes, rather than other audiences like market researchers or other commercial uses.

### Find your Facebook ID:

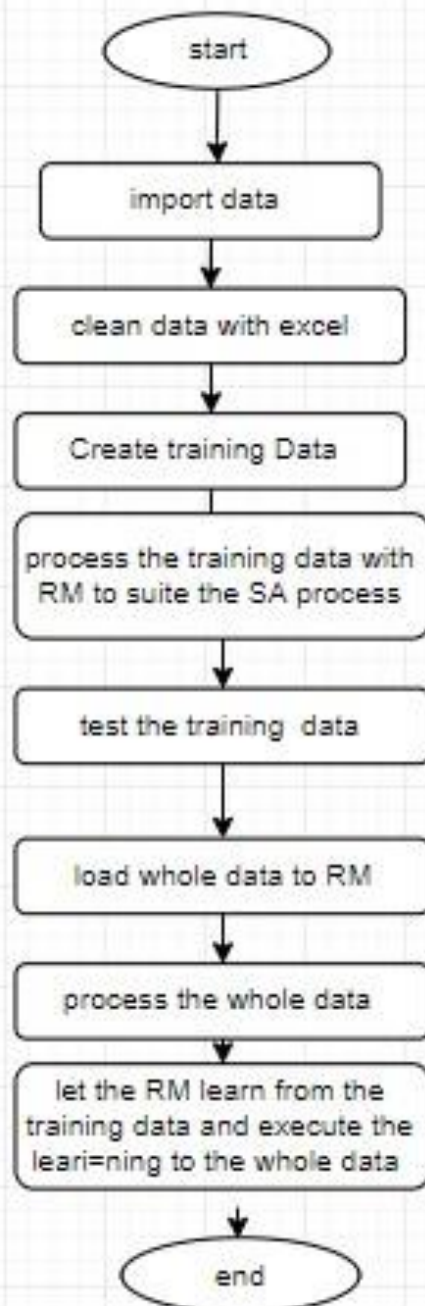
Tool turn our Url in to numerical ID

#### Find your Facebook ID

To find your Facebook personal numeric ID for fb:admins, social plugins, and more, enter your Facebook personal profile URL below:

Find numeric ID →

**And this is a simplest flow chart that summarized our work:**



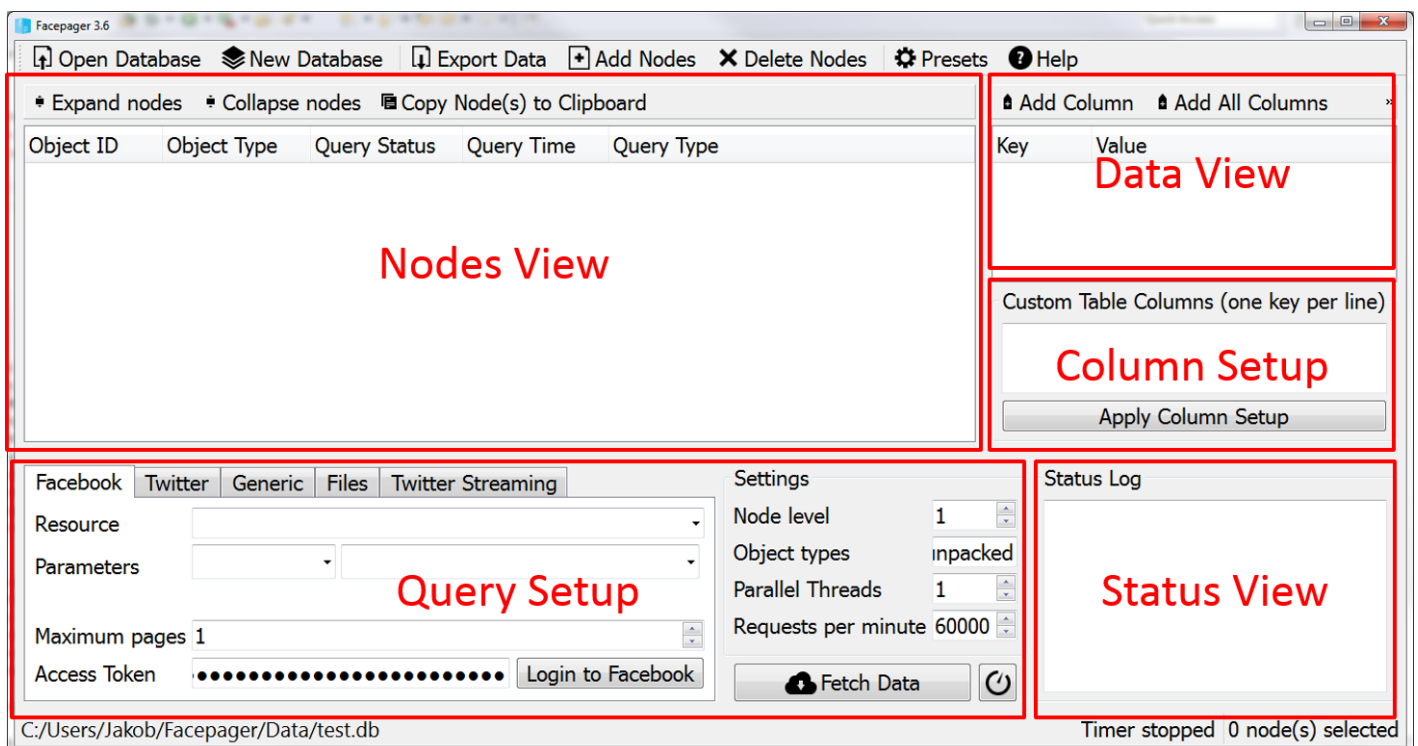


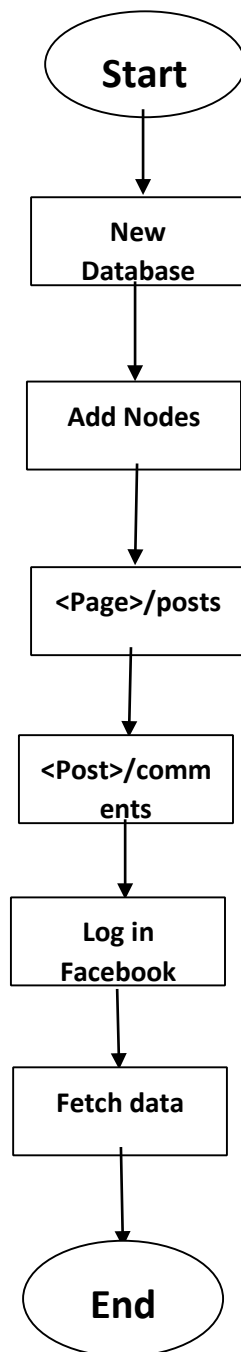
# ETL (Extract, Transform, load) For Facebook Data

## Facepager5.3.

We use the Facepager tool to import the data from Facebook.

**Facepager** was made for fetching public available data from Facebook, twitter, YouTube and other JSON-based API. All data is stored in a SQLite database and maybe exported to csv.





And the data we exported is shown as below:

This is the final result in the csv file.

After we export the data in the excel files we notice that it is noisy as shown below:





And we find that we can process its form into a better one that the Rapid miner tool can deal with, **First:**

We must remain just the comments column

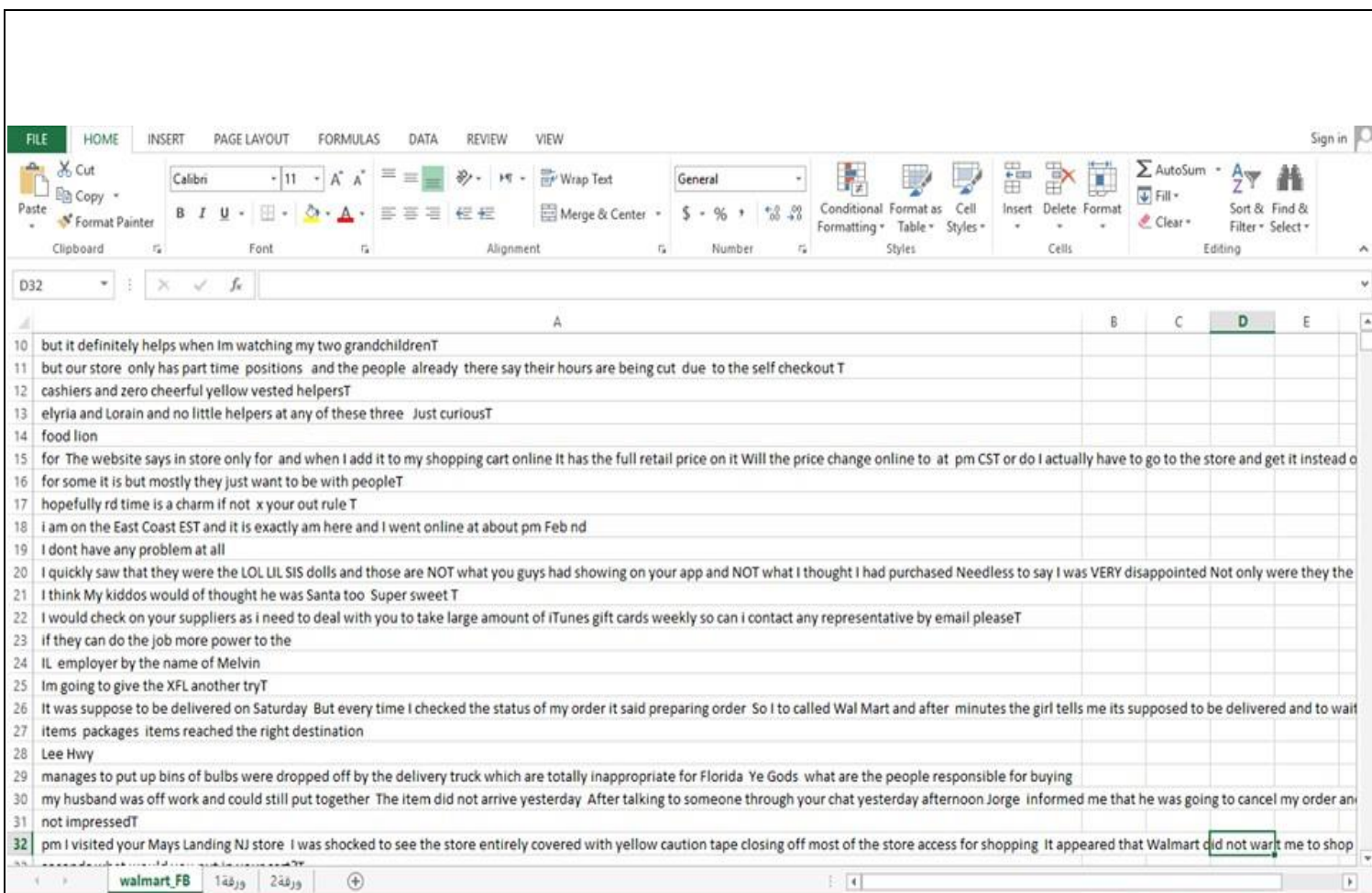
Microsoft Excel interface showing a spreadsheet with columns A1 through Z.

The spreadsheet contains a single row of data starting from column A:

	A
1	"81;\"2;\", \"10156306969234236_10156309298519236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.546000\",\"Facebook<post>/comments\", \"Walmart is the greatest place on earth
2	"82;\"2;\", \"10156306969234236_10156307640594236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.546000\",\"Facebook<post>/comments\", \"I probably gonna to get that for my kids\", \"2018-02-05T21:21:21+00:00
3	"83;\"2;\", \"10156306969234236_10156307412574236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Anyone out there help me understand the new Savings Catcher
4	"84;\"2;\", \"10156306969234236_10156307264724236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Good luck getting your items if Walmart ships them via Laser
5	"85;\"2;\", \"10156306969234236_10156308064384236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Can Walmart stores set up their own return policies separate
6	"86;\"2;\", \"10156306969234236_10156307847624236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"I had a customer service experience 7535 South Ashland
7	"87;\"2;\", \"10156306969234236_10156307158964236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Good treat for Valentine's Day\", \"2018-02-05T17:16:25+00:00
8	"88;\"2;\", \"10156306969234236_10156312935859236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"I have always been a big fan of the Walmart Savings Catcher
9	"89;\"2;\", \"10156306969234236_10156310372224236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"I was an employee for a very short time. I was called into the
10	"90;\"2;\", \"10156306969234236_10156310291279236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"don't know if this is the correct place to try and get my problem
11	"91;\"2;\", \"10156306969234236_10156309156894236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Walmart in Wooster Ohio is now a nightmare and I dread going
12	"92;\"2;\", \"10156306969234236_10156311618769236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"put the Lego anniversary bundle in cart at 8 this morning
13	"93;\"2;\", \"10156306969234236_10156309226549236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"I've been a long-time Walmart shopper and card holder. I never
14	"94;\"2;\", \"10156306969234236_10156308128009236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Are you guys going to fix the careers page? tried logging in and
15	"95;\"2;\", \"10156306969234236_10156310190394236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"I am having a horrible time with the care plan I purchased with
16	"96;\"2;\", \"10156306969234236_10156310099184236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Since there is no way to place a regular comment I guess i'll p
17	"97;\"2;\", \"10156306969234236_10156315810514236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Worst experience ever at Watertown
18	"98;\"2;\", \"10156306969234236_10156308929919236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"As I cannot send a message or post to page: have you seen t
19	"99;\"2;\", \"10156306969234236_10156309244744236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"DONT shop on walmart.com...I made a purchase and the deli
20	"100;\"2;\", \"10156306969234236_10156314644934236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"How do I print off my savings catcher now that it has change
21	"101;\"2;\", \"10156306969234236_10156311492349236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"I was chatting with Paul G on live assistances and as I was l
22	"102;\"2;\", \"10156306969234236_10156313350689236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Yo Mac. after standing around the deli for five minutes why
23	"103;\"2;\", \"10156306969234236_10156312724809236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Happy Valentine's Day Walmart. Just wanted to let you know
24	"104;\"2;\", \"10156306969234236_10156311641964236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"Stood at the Deli counter in the Elk Grove store two different
25	"105;\"2;\", \"10156306969234236_10156310291279236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"I don't know if this is the correct place to try and get my prob
26	"106;\"2;\", \"10156306969234236_10156310096649236\",\"data\",\"fetched (200)\",\"2018-02-09 19:07:44.878000\",\"Facebook<post>/comments\", \"I just tweeted photos to the Walmart and Lay's twitter accoun

## Second:

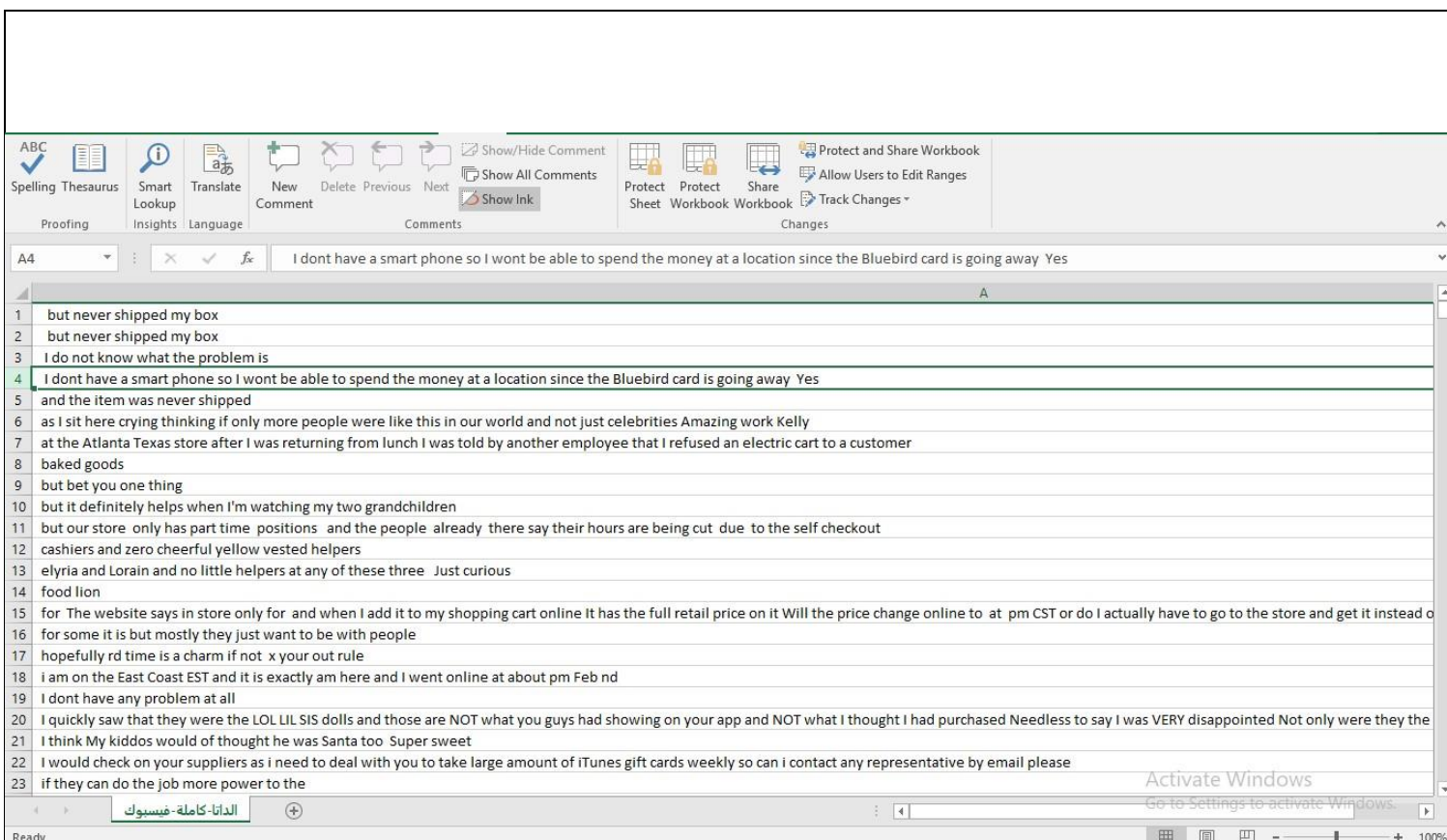
We found the beginning of each comment contain the same size of unusual data so we remove it using the "mid" function in excel for them all at the same time as "MID (text,128, (LEN(text)-128))" and the result is as shown below:



After that we found that the most of comments had a "T" letter in the end so we decide **Third:**

to remove it also with excel in a quick useful way using the "if" condition and the "mid" function as `"if(RIGHT(text)="T", MID(text,1, LEN(text)-1), text)"` and the result is as shown below:

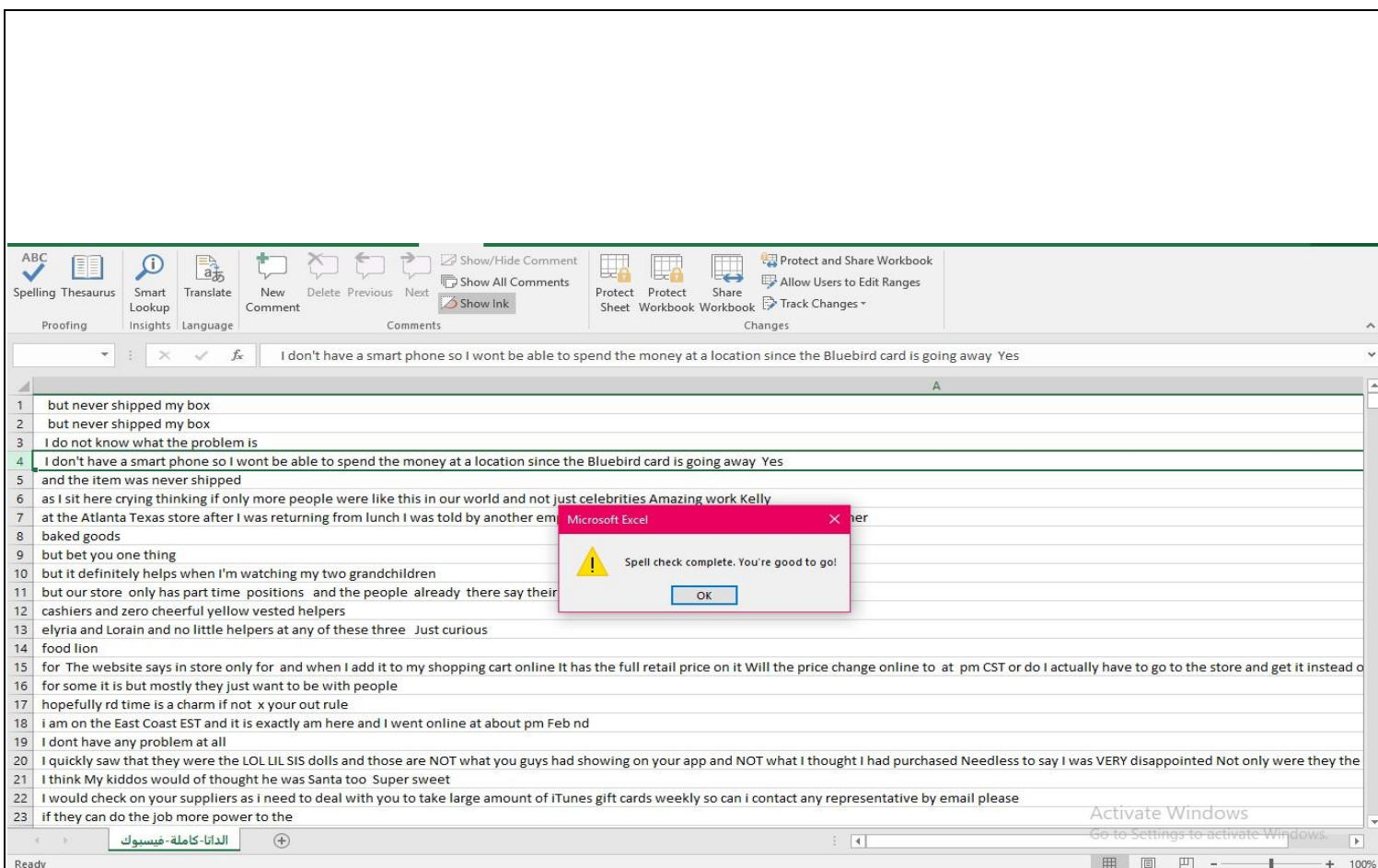




Here we notice that there is some of words written in a wrong form, then we search about this problem and find that we can solve it by using the **spelling process** (Corrects how the word is typed) in excel as the following:

For example, as shown in the picture above the word “don’t” is not write in the correct way, also in line ‘13’ the beginning of the word must be in capital letter “elyria” and so on in the remaining data, spelling process can solve these problems,

Then the list shows the wrong word and suggestions to choose the correct one. We click to the correct word “don’t” and click change to change it to the correct form as shown:

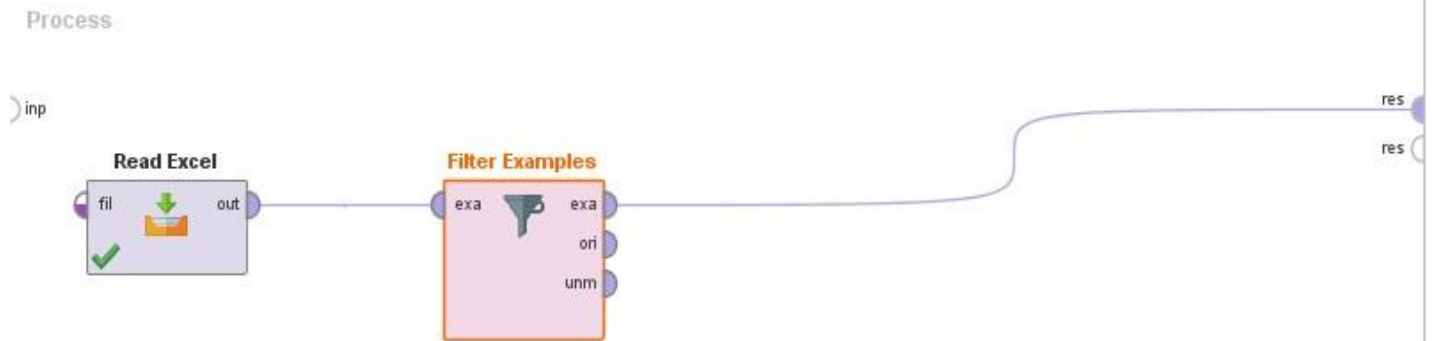


**“spell check complete. You’re good to go!”**

The same we are used to correct the remaining data.



The data is cleaned, but we want to remove the empty rows using the “filter example” process in rapid miner:



**Here the data is cleaned enough to enter the processes of rapid Mainer.**

## The creation of training data:

In the beginning, after understanding and studying data, we work on training data on the basis that the data relating to the services provided by Walmart such as customer service, technical support, online purchasing, employee handling, without looking at the data that belong to the products sold by Walmart and not produced by himself, so that there is transparency and credibility in the process of analysis of data that belong to this type of work, the data was taken randomly and then categorize them to positive, negative and neutral manually, we tried to some extent for data to be equally in the three classes(positive,negative,neutral), The data was split between the project members and each one of us did a piece of work, study, analyze and classify it. Then we performed a review of each part by the rest of the team members to verify the validity of the classification. As for the number of data selected on the basis of the experiment, we try to experiment more than a number and we apply the model and finally we have stabilized on the number that gives the highest accuracy.

## Load data by Rapidminer:

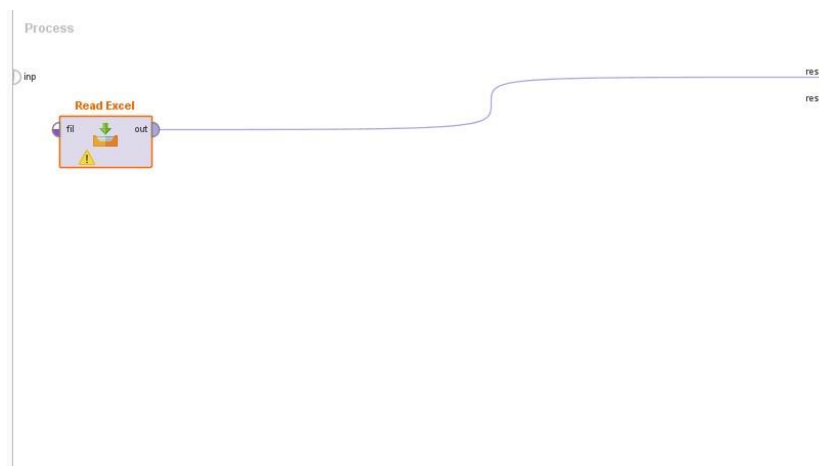
Sample of train data that we use to major the accuracy, in first step we use to class to determine if the comment is positive or negative.

A		B
1	text	result
2	Walmart very nice	positive
3	I love shopping at Walmart	positive
4	you came to their aid Thank you for what you do Walmart	positive
5	I always receive friendly and helpful service from the Walmart in Marinette	positive
6	We went and enjoyed the treats at the Perrysburg	positive
7	Walmart is the worst	negative
8	worst customer service ever Walmart online chat wont help me	negative
9	Walmart provides the WORST customer service and with each new change in their return policy it results in them stealing from you	negative
10	Worst experience ever at Walmart	negative
11	Hate Walmart	negative
12	Walmart does not like my comment because it was a bad experience so they marked it as spam	negative

## First:

we use **Read Excel** operator:

This operator reads an Example Set (sample data) from the specified Excel file, our excel file contain the train data we use, we use import configuration wizard for loading data from excel file to rapid miner Here we see that our excel file is upload on the tool, if we click run we can check that and see our data



## Secondly:

**Set role:** This Operator is used to change the role of one or more Attributes.

### Parameters

**Attribute name:** the name of the Attribute which role should be changed. The name can be selected from the dropdown menu or manual typed. (result as we name in our train file).

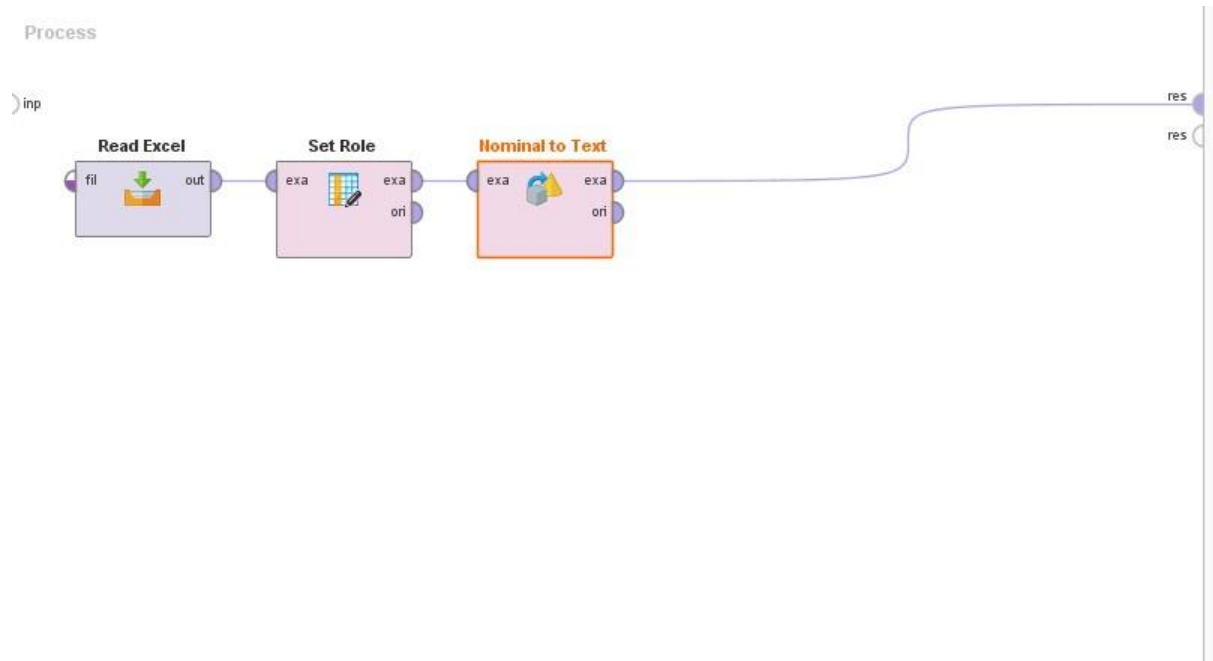
**Target role:** The target role of the selected Attribute is the new role assigned to it.

**Label:** This is a special role. An Attribute with the label role acts as a target Attribute for learning Operators. The label is also often called 'target variable' or 'class'.



## Thirdly:

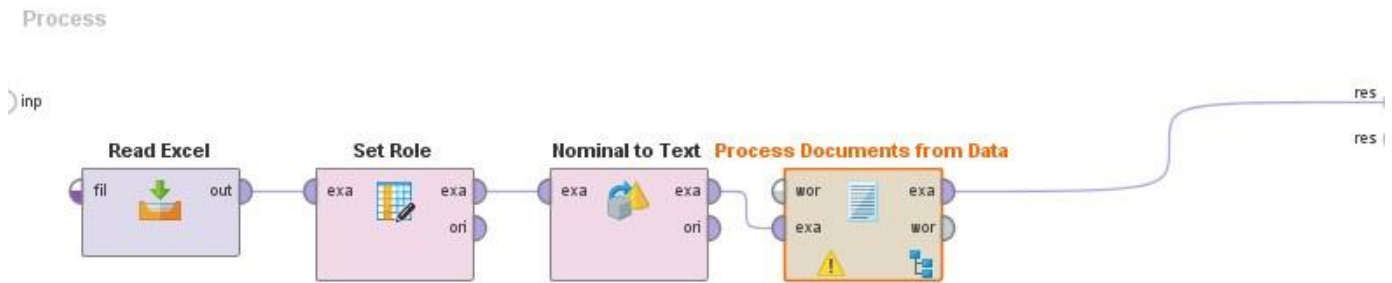
**Nominal to text** (To specify which column is a text column, since Rapidminer "Process Documents..." Operators work only on text data.



## Fourthly:

We use the **Process documents from data** operator is used to create word vectors from text attributes, this is a nested operator that contain sub operators inside it.

**TF-IDF** stands for term frequency–inverse document frequency. It is a numerical statistic which reflects how important a word is to a document in a collection, and it is often used as a weighting factor.



The **Tokenize** operator tokenizes documents, and we select in the parameters of this operator to tokenize at non letters so that each time a non-letter is found it shall denote a new token, therefore splitting a text into words (This operator splits the text of a document into a sequence of tokens)

The **Filter Stopwords (Dictionary)** operator applies a stopwords list from a file.

Stopwords are words which are filtered out prior to, or after, processing of natural language data (text)

e. For example, some of the most common stopwords for search machines are: the, is, at, which, and on.

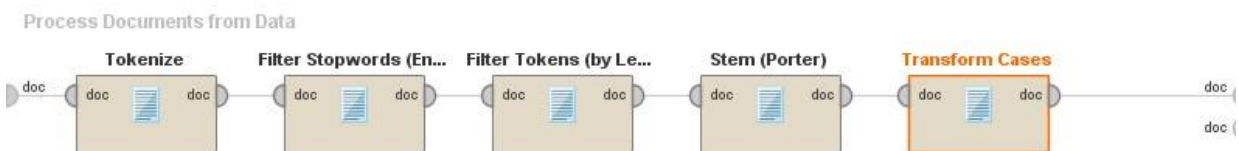
) Words that do not matter when parsing text(

The **Filter Tokens (by Length)** operator, filters tokens based on their length. In its parameters we select the min chars of a token to be 3 (thus

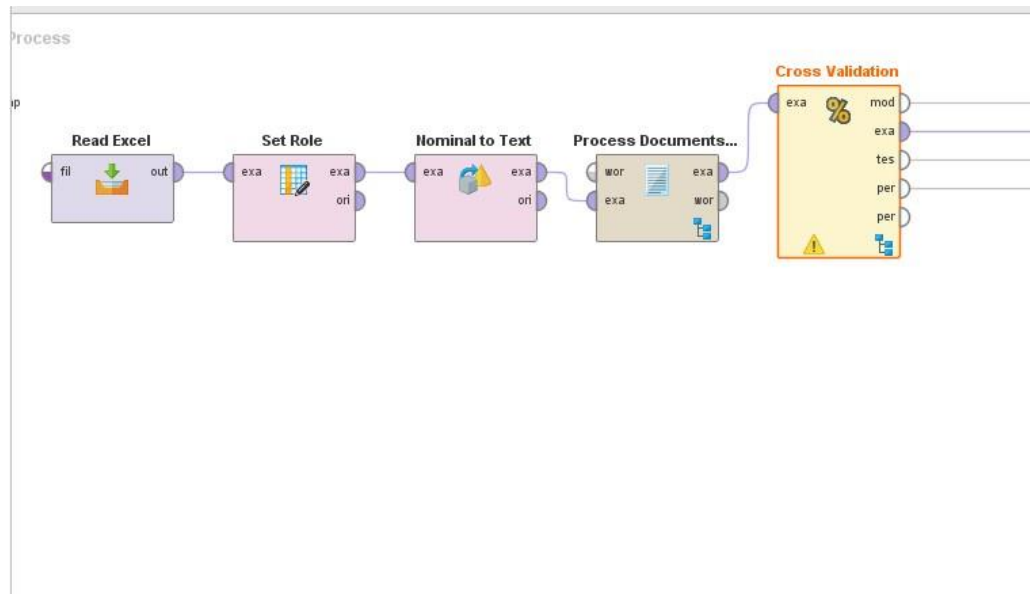
removing single letter words), and the max chars of a token to be 20 which is safe enough to say that words consisting of 20 chars are probably gibberish.

Stemming also known as lemmatization is a technique for the reduction of words into their stems, base or root. Many words in the English language can be reduced to their base form or stem e.g. like, liking, likely, unlike belong to like.

The Transform Cases operator transforms the cases of all characters. In its parameters we choose to transform all characters to lower case.



**Sixth:** we use the cross validation process as shown below:



Test	Train	Train	Train	Train	Train	Train	Train	Train	Train
------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Max -> best validation

Max->best learning result

Number of train data sets= 500 row (comment, class)

K (number of folds) =10, number of rows (data set size) =500

$500/10= 50$

50	50	50	50	50
50	50	50	50	50

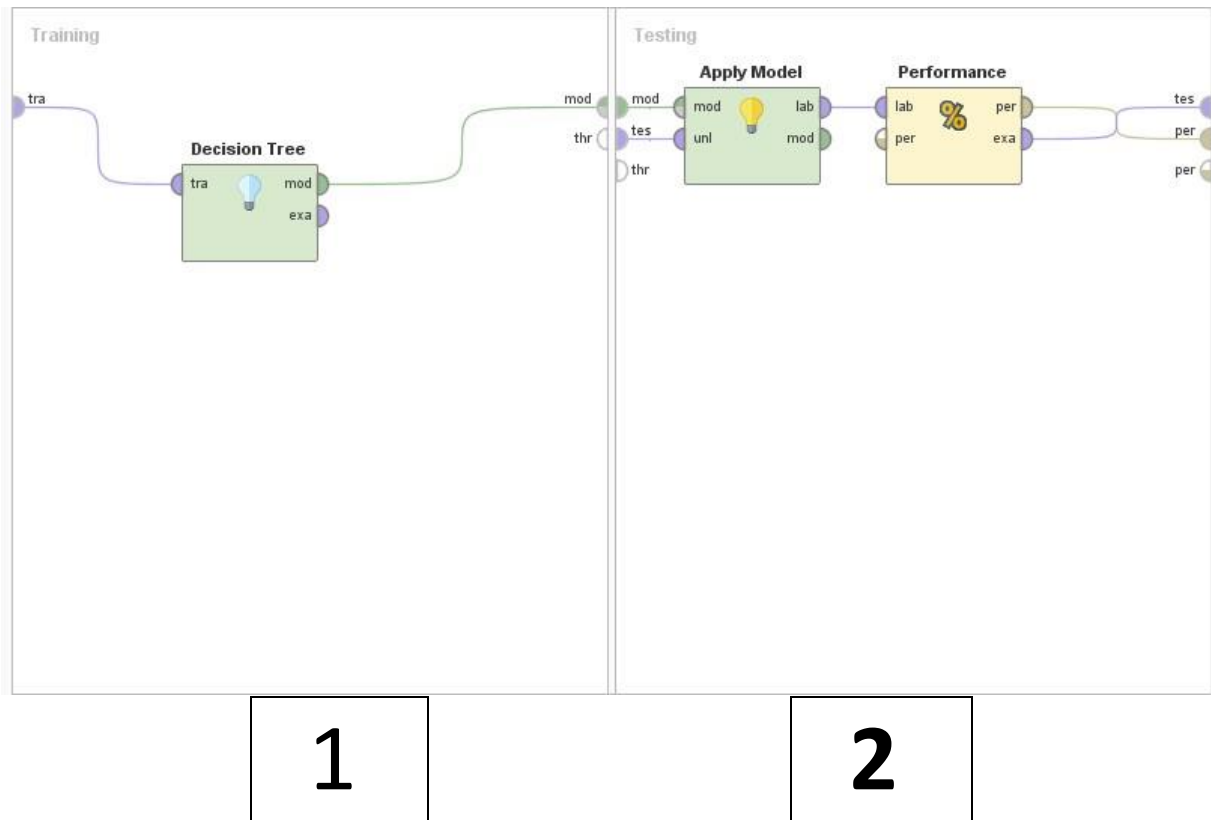


The Cross Validation Operator is a nested Operator. It has two sub processes: A Training sub process and a Testing sub process. The Training sub process is used for training a model. The trained model is then applied in the Testing sub process. The performance of the model is measured during the Testing phase.

In the beginning, cross validation is a process that divides train data into parts,  $K$  is chosen as 10, and according to many experiments for more numbers we discover that number 10 is the one that gives the highest accuracy. In addition, we investigated several files of the subject and the majority advice to choose the number 10 to give better results and better learning ability. **cross validation works as follows:**

Divide the 10 groups into two parts as shown in the table above. The first section contains  $(9/10)$  groups which contains the part that the model will learn from, and the second section contains one set  $(1/10)$  which is the part that we will do a test on it and show us a certain accuracy, after that this process will make reverses groups that take 9 different groups for learning and one group for testing process and returns to calculate the accuracy in the same way and so on (the process does the switching process until it finishes all the groups (train, test) and takes the average of All the accuracy that appeared in all operations, and shown to us as a final result rate and this is the result on which to calculate whether the accuracy is high or not.

And this process contains sub processes as shown below:




**Number 1 is the (training part)** that contains the Classification model (decision tree).

**Number 2 is the (testing part)** that contains: -

1- The apply model (This Operator applies a model on an Example Set (train set)).

2- The performance operator (This operator is used for performance evaluation. It delivers a list of performance criteria values. These performance criteria are automatically determined in order to fit the learning task type.)

After We click on RUN () we can notice the program run by see the cross validation operator

accuracy: 83.52% +/- 4.86% (mikro: 83.52%)

	true positive	true nigtive	class precision
pred. positive	204	23	89.87%
pred. nigtive	66	247	78.91%
class recall	75.56%	91.48%	

Through analyzing comments (positive, negative), we discovered a comments that is not positive and not negative(neutral), so we decide to add a new class called: **neutral**.

Here is a sample of train data that contains three classes (positive, negative, neutral) that we use it to measure the accuracy, to ensure that the machine learning is applied on training data (pos, neg, nut) in the correct way or not (does the model learn right from the human or not?).

text	result
so well organized Thankful for all of the great employees	positive
which is great for me because I have a baby Thanks Walmart You are doing a great job	positive
you came to their aid Thank you for what you do Walmart	positive
I love shopping at WalMart even when crowded especially toys and books and this picture is awesome	positive
I love the Walmart where I live They have everything	positive
Walmart is the greatest place on earth	positive
WalMart Really not cool	negative
Walmart is the worst	negative
WORST EXPERIENCE EVER THANKS WALMART from a former customer	negative
After waiting an eternity for my online purchase and bad treatment for their customer service	negative
I was My local walmart had a very very bad experience on January	negative
but Walmart has lost some business here Very disappointed in the store and service	negative
Walmart online is useless	negative
shop at Walmart	nutral
Walmart sets the standard In this case	nutral
Walmart im trying to buy a gift set that is in your black friday ad but it shows a different price Can you help me please Pease	nutral
Do Walmart have whatsapp group chat	nutral
Hello I was wondering if Walmart does international shipping to Argentina	nutral
Walmart I Was wondering if you guys carry security jackets?	nutral

After that we build the model (with different operators) that described in details in the previous pages (in training data for 2 classes), and **run** the process to show the results below:

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results

Result History x PerformanceVector (Performance) x

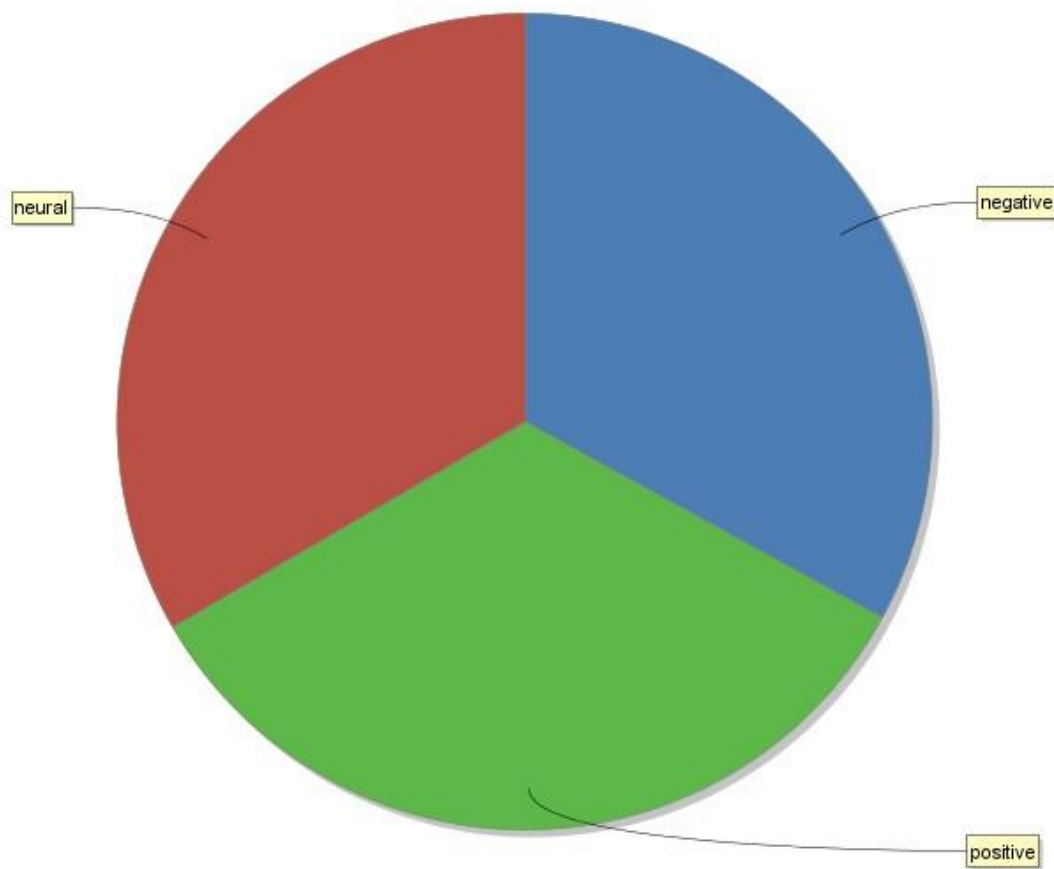
ExampleSet (Cross Validation) x ExampleSet (Cross Validation) x Tree (Decision Tree) x

ExampleSet (809 examples, 5 special attributes, 1777 regular attributes) Filter (809 / 809 examples): all

Row No.	positive	prediction(p...	confidence(nutral)	confidence(positive )	confidence{...	aai	aand	abl	abs
1	positive	nutral	0.462	0.077	0.460	0	0	0	0
2	positive	positive	0.085	0.872	0.043	0	0	0	0
3	positive	nutral	0.462	0.077	0.460	0	0	0	0
4	positive	positive	0	1	0	0	0	0	0
5	positive	positive	0	1	0	0	0	0	0
6	positive	positive	0	1	0	0	0	0	0
7	positive	positive	0.085	0.872	0.043	0	0	0	0
8	positive	nutral	0.462	0.077	0.460	0	0	0	0
9	positive	positive	0	1	0	0	0	0	0
10	positive	positive	0.085	0.872	0.043	0	0	0	0
11	positive	positive	0	1	0	0	0	0	0
12	positive	nutral	0.462	0.077	0.460	0	0	0	0
13	positive	positive	0	1	0	0	0	0	0
14	positive	positive	0	0.986	0.014	0	0	0	0
15	positive	positive	0	0.986	0.014	0	0	0	0

Activate Wi Go to Settings

Here the picture shows the **positive column** (training data) and **prediction column**) the result predicted by the model), and the percent of confidence for each class. For example, in the first row the training data is (positive) and the prediction is (neutral), this means that the model is not learning well from the training data) the result that the model predict is wrong).



accuracy: 56.99% +/- 3.33% (mikro: 56.98%)

	true positive	true negative	true nutral	class precision
pred. positive	198	17	7	89.19%
pred. nignative	0	0	0	0.00%
pred. nutral	71	253	263	44.80%
class recall	73.61%	0.00%	97.41%	

The accuracy from training data in 3 classes (positive, negative, neutral) is 56.99 which is not good enough, so we want to use a new way that increase the accuracy and that make the model distinguish between classes and learn well from training data. (the details about this in the next pages).

As we see above the accuracy of the learning is not good enough so we found that we should increase it by **First:**

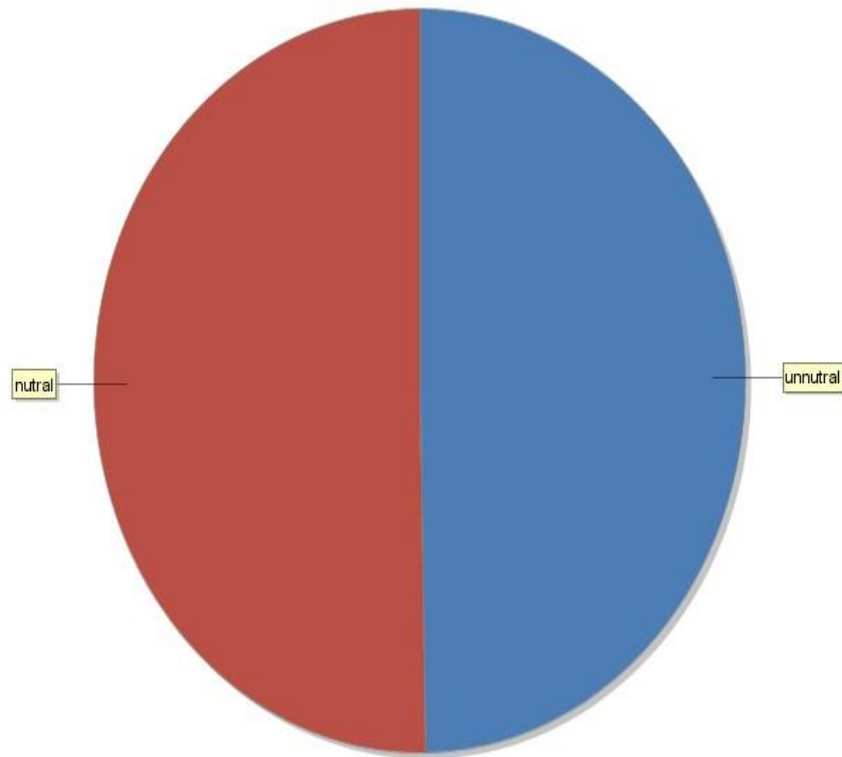
Test the model with two classes (Neutral, Unnatural) and measure the accuracy.

## Second:

Test the model with another two classes (positive, negative) and measure the accuracy and then compute the average between two measure accuracy as we will see below:

1	text	result
2	Walmart is the worst	unnutral
3	WalMart Really not cool	unnutral
4	I hate your page I wanted the phone number for my walmart couldnt get it the store finder has no button to push when you put your city and state	unnutral
5	The staff at the new Walmart in my area arent helpful at all They displayed how much they hate working there all over their faces Please correct that before too late	unnutral
6	have a very bad experience at Fairfax Walmart Supercenter	unnutral
7	Andrea did you look in the outdoor section? At my Walmart they had the Holiday CD display by the blow up Christmas stuff in the Home and Garden Section	nutral
8	Walmart in roxboro NC will take them	nutral
9	Walmart dont know when theyll get them in again	nutral
10	hello walmart how can i change my shipping address	nutral
11	When I click on refill prescription in My Walmart pharmacy account	nutral

Using **First** case with (Neutral, Unneutral)-Training Data:



We use this model –that is explained previously– to compute the accuracy of the previous training data, and we find notice that the accuracy is 76.75 as we see here:

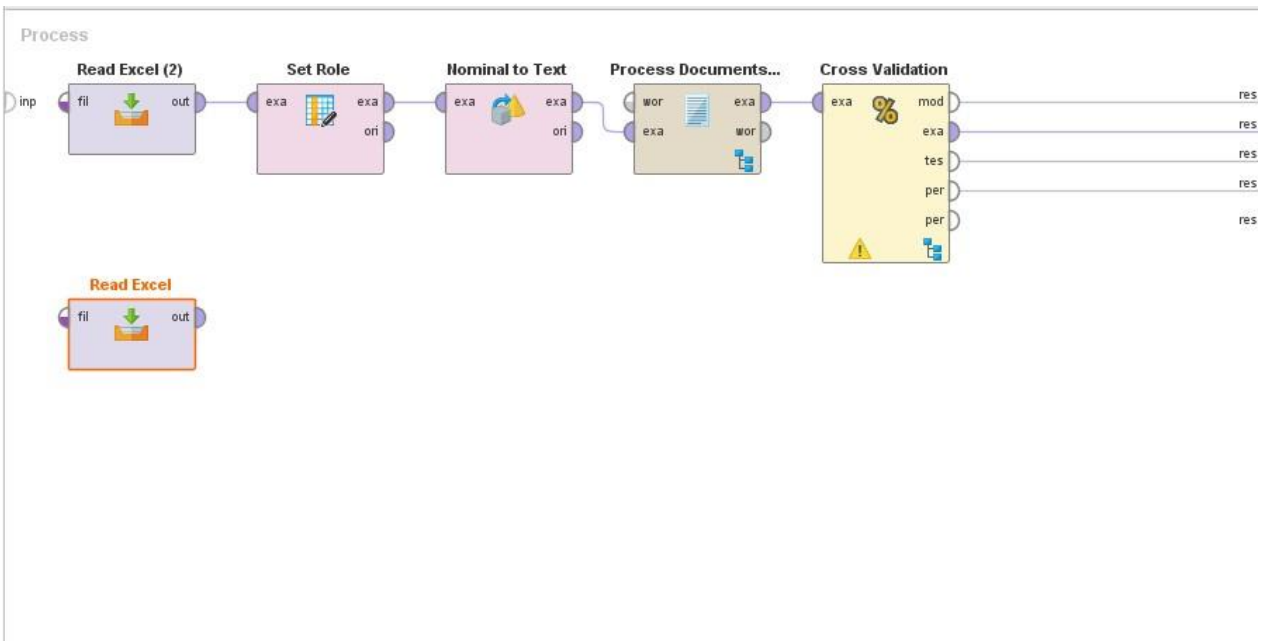
accuracy: 76.75% +/- 4.60% (mikro: 76.72%)

	true unneutral	true nutral	class precision
pred. unneutral	157	15	91.28%
pred. nutral	110	255	69.86%
class recall	58.80%	94.44%	

Here we will execute the classification as (neutral and unneutral) to whole Walmart Facebook data so:



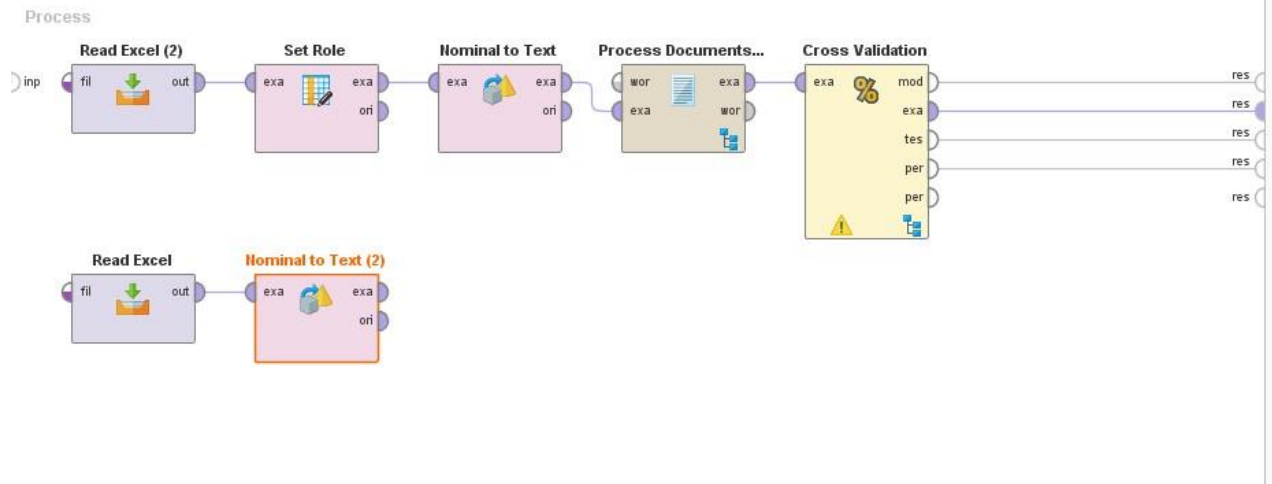
**First:** we use the read excel process to load all Walmart Facebook data in rapidminer as shown below:



**Second:**

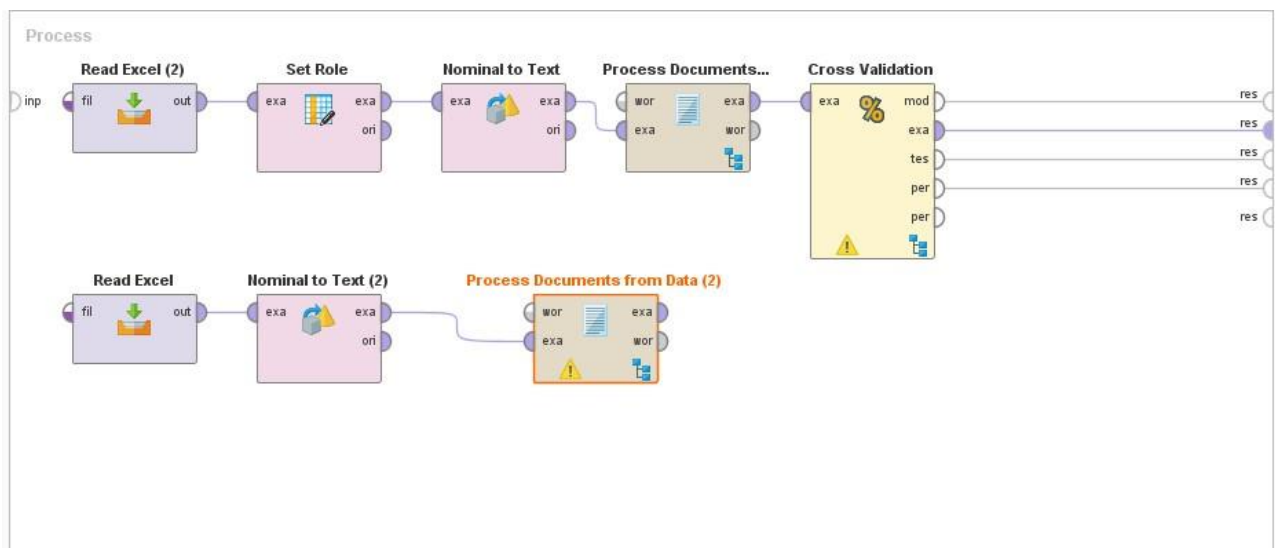
we use the Nominal to text process again as we made in the training data (To specify which column is a text column, since Rapidminer "Process Documents..." Operators -that will be used in the next process - work only on text data.

As we will see below:



## Third:

we used the (Process documents from data) process that we explain it previously to clean the text using a lot of sub processes inside of it as we will see below:



And this process as previously mentioned it contain a lot of sub processes as we see below:



The **Tokenize operator** tokenizes documents, and we select in the parameters of this operator to tokenize at non letters so that each time a non-letter is found it shall denote a new token, therefore splitting a text into words (This operator splits the text of a document into a sequence of tokens)

The **Filter Stopwords(Dictionary) operator** applies a stopwords list from a file. Stopwords are words which are filtered out prior to, or after, processing of natural language data (text)

e. For example, some of the most common stopwords for search machines are: the, is, at, which, and on.

) Words that do not matter when parsing text(

The **Filter Tokens (by Length) operator**, filters tokens based on their length. In its parameters we select the min chars of a token to be 3 (thus removing single letter words), and the max chars of a token to be 20

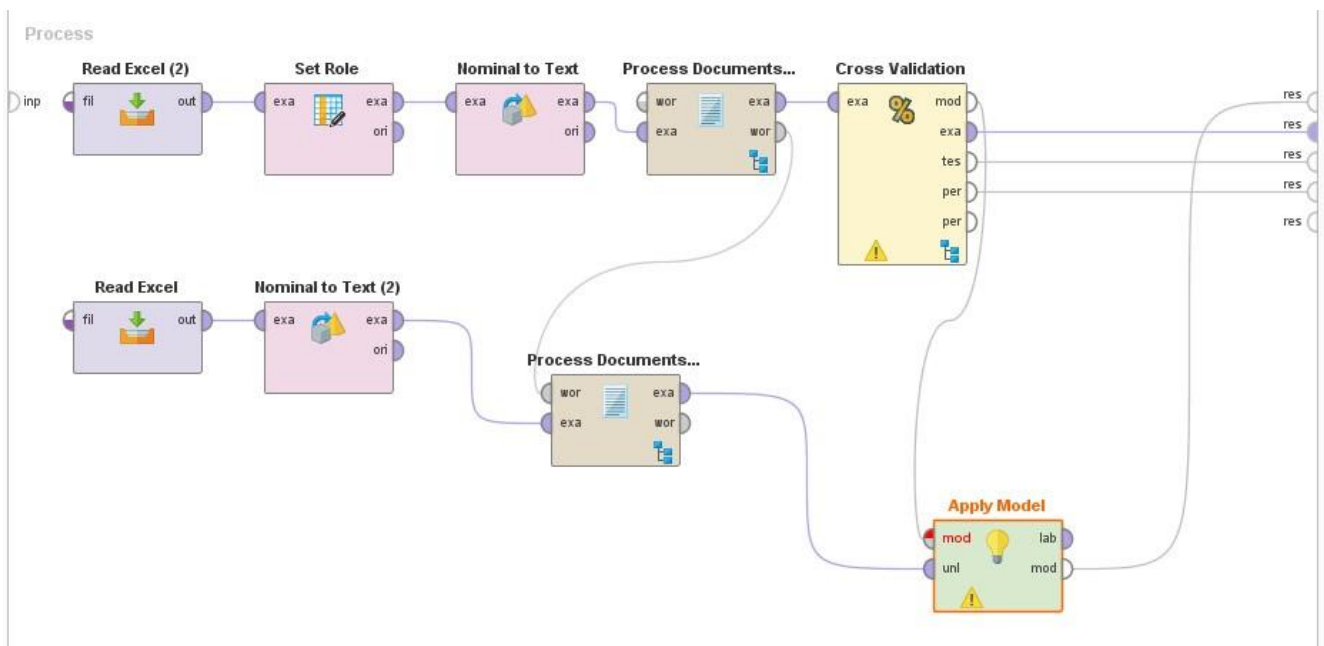
which is safe enough to say that words consisting of 20 chars are probably gibberish.

**Stemming** also known as lemmatization is a technique for the reduction of words into their stems, base or root. Many words in the English language can be reduced to their base form or stem e.g. like, liking, likely, unlike belong to like.

The **Transform Cases** operator transforms the cases of all characters. In its parameters we choose to transform all characters to lower case

## Forth:

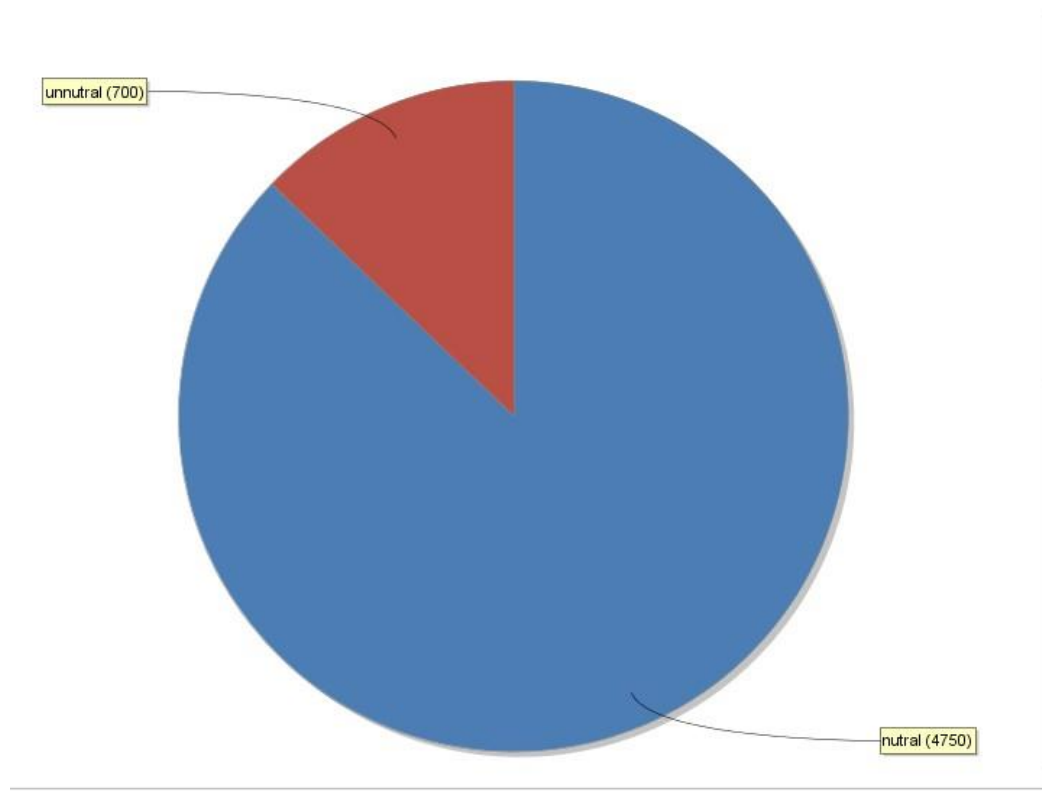
we execute the model using the (Apply Model) process, This Operator applies a model on an Example Set. As shown below:











And **secondly** after we extract the neutral and unneutral data we take all unneutral data and test it using this training data (positive, negative)-training data:

A		B
1	text	result
2	Walmart very nice	positive
3	I love shopping at Walmart	positive
4	you came to their aid Thank you for what you do Walmart	positive
5	I always receive friendly and helpful service from the Walmart in Marinette	positive
6	We went and enjoyed the treats at the Perrysburg	positive
7	Walmart is the worst	negative
8	worst customer service ever Walmart online chat wont help me	negative
9	Walmart provides the WORST customer service and with each new change in their return policy it results in them stealing from you	negative
10	Worst experience ever at Walmart	negative
11	Hate Walmart	negative
12	Walmart does not like my comment because it was a bad experience so they marked it as spam	negative



and using the previous model we find the accuracy 83.31 as shown below

accuracy: 83.31% +/- 4.67% (mikro: 83.30%)

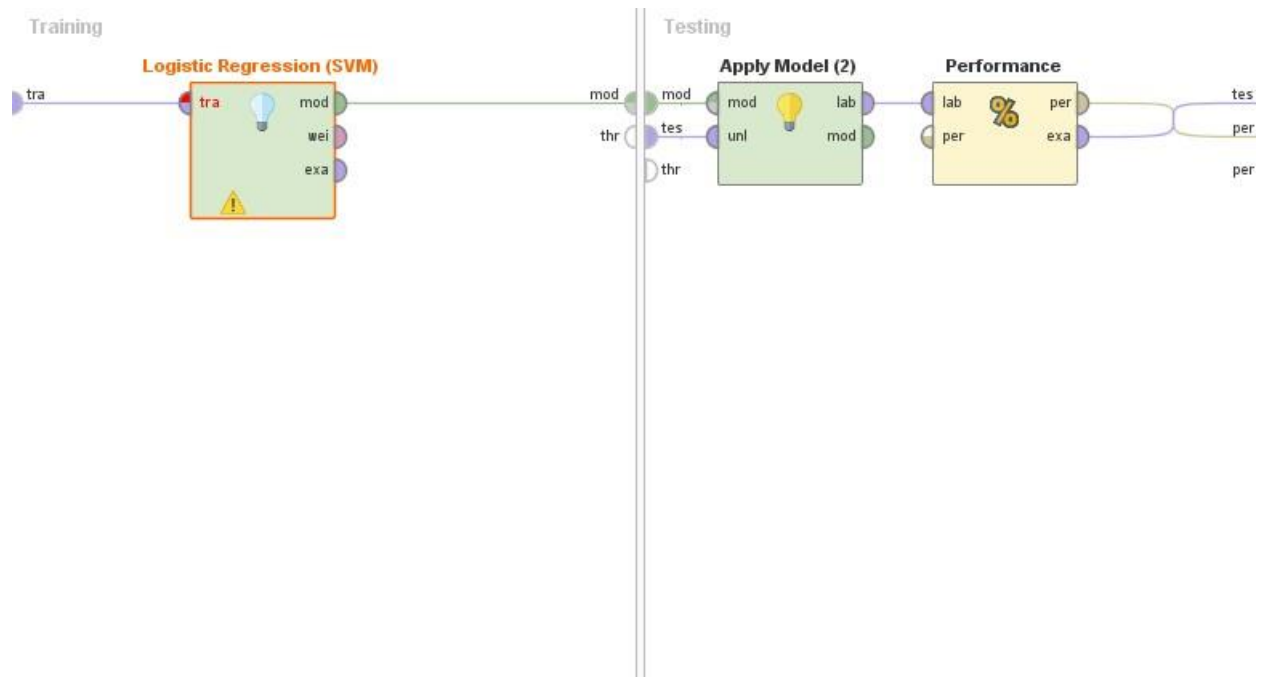
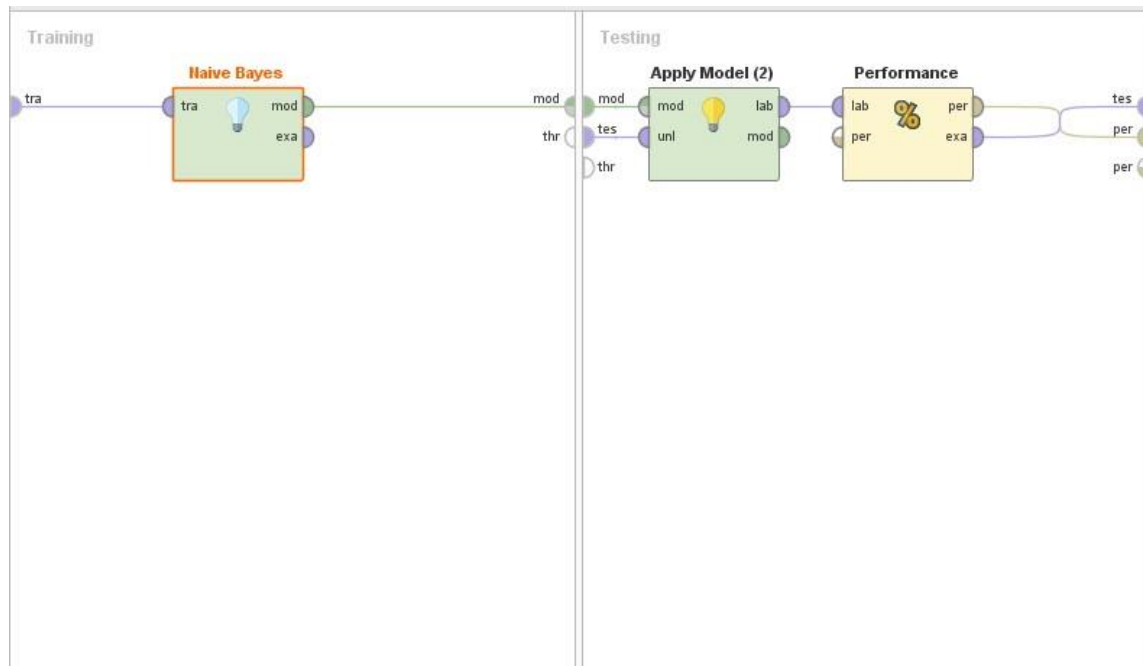
	true positive	true negative	class precision
pred. positive	201	22	90.13%
pred. negative	68	248	78.48%
class recall	74.72%	91.85%	

Here we found that the accuracy of using these two processes is  
 $(76.62 + 83.31) / 2$

$$= 79.9$$

This result is better than using (positive, negative, neutral) as a training data

And we also we try it using deferent classification models like (SVM, naïve bays) as follows:

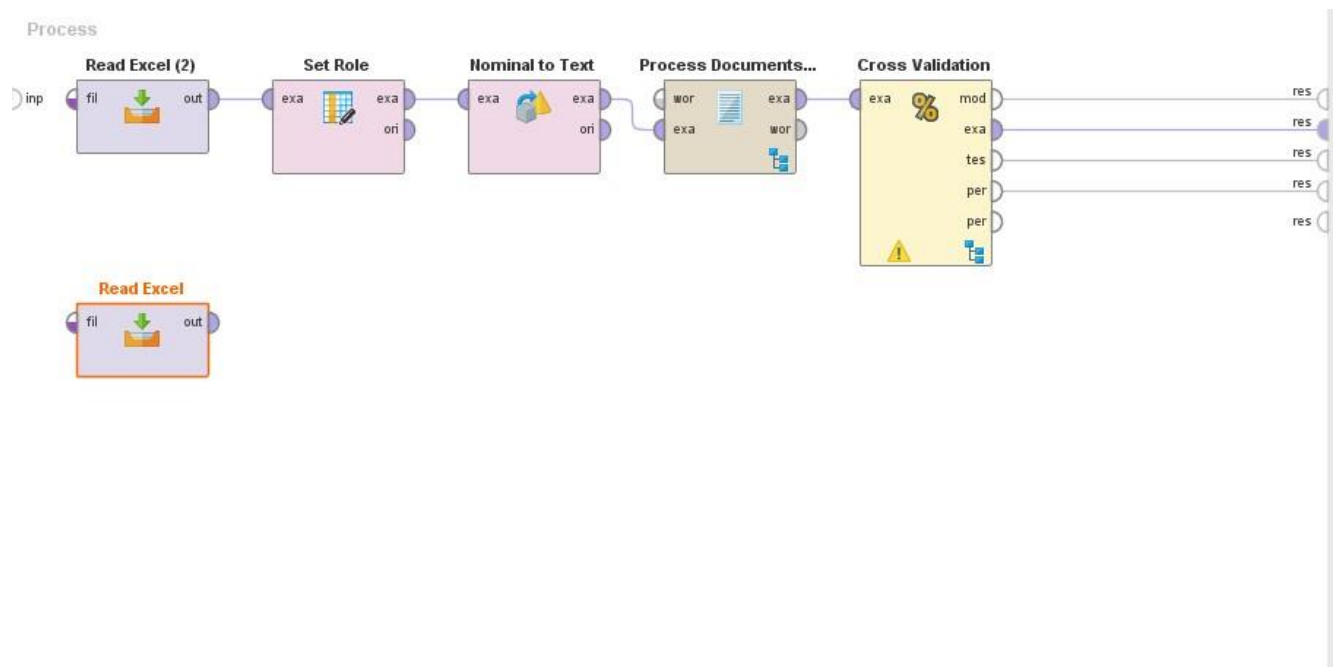


And the comparison between them will appear in the result page

So we will complete our model with this case of training data

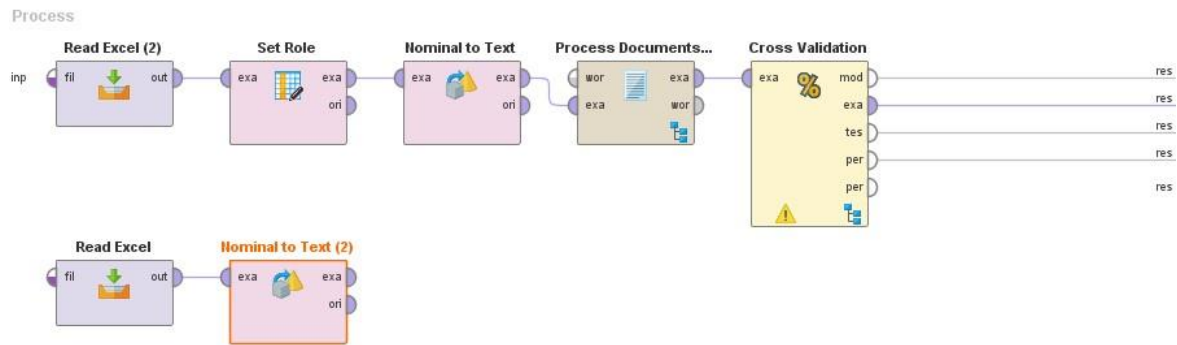
Let's start to building our model using the training data that we reached in previous page and make the prediction to the whole Walmart Facebook data.

**First:** we use the read excel process to load all Walmart Facebook data in rapidminer as shown below:

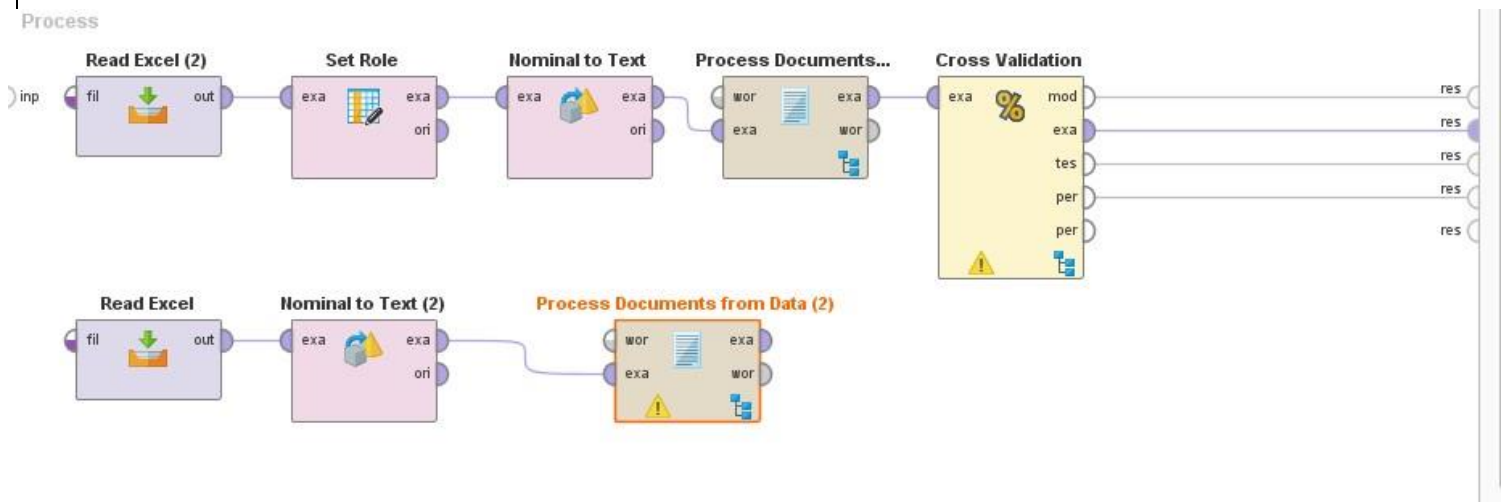


**Second:** we use the Nominal to text process again as we made in the training data (To specify which column is a text column, since Rapidminer "Process Documents..." Operators -that will be used in the next process - work only on text data.

As we will see below:

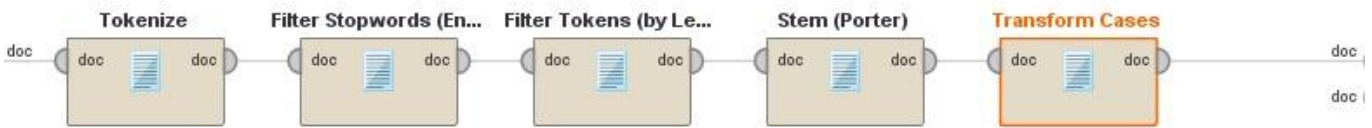


**Third.** we used the (Process documents from data) process that we explain it previously to clean the text using a lot of sub processes inside of it as we will see below:



And this process as previously mentioned it contain a lot of sub processes as we see below:

#### Process Documents from Data



The **Tokenize operator** tokenizes documents, and we select in the parameters of this operator to tokenize at non letters so that each time a non-letter is found it shall denote a new token, therefore splitting a text into words (This operator splits the text of a document into a sequence of tokens)

The **Filter Stopwords(Dictionary) operator** applies a stopwords list from a file. Stopwords are words which are filtered out prior to, or after, processing of natural language data (text)

e. For example, some of the most common stopwords for search machines are: the, is, at, which, and on.

) Words that do not matter when parsing text(

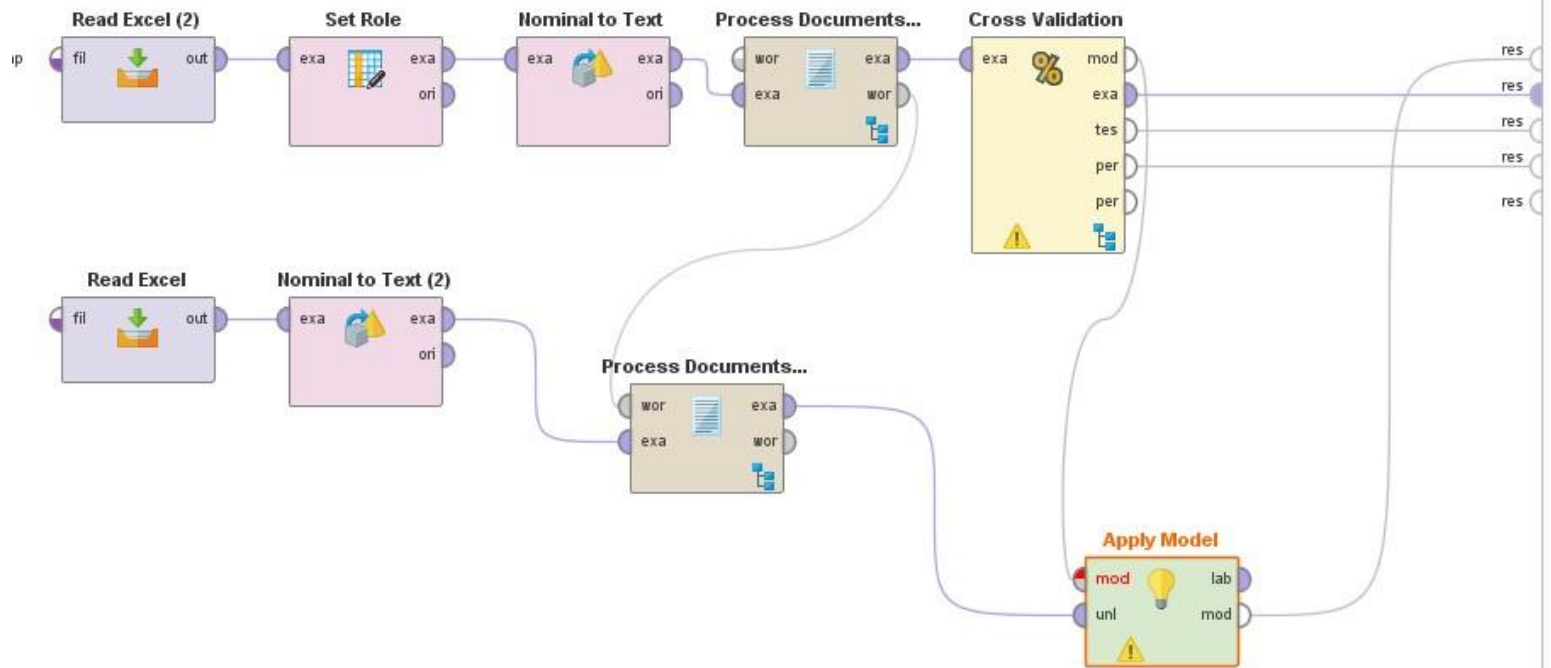
The **Filter Tokens (by Length)** operator, filters tokens based on their length. In its parameters we select the min chars of a token to be 3 (thus removing single letter words), and the max chars of a token to be 20 which is safe enough to say that words consisting of 20 chars are probably gibberish.

**Stemming** also known as lemmatization is a technique for the reduction of words into their stems, base or root. Many words in the English language can be reduced to their base form or stem e.g. like, liking, likely, unlike belong to like.

The **Transform Cases** operator transforms the cases of all characters. In its parameters we choose to transform all characters to lower case.

**Forth:** we execute the model using the (Apply Model) process, This Operator applies a model on an Example Set. As shown below:

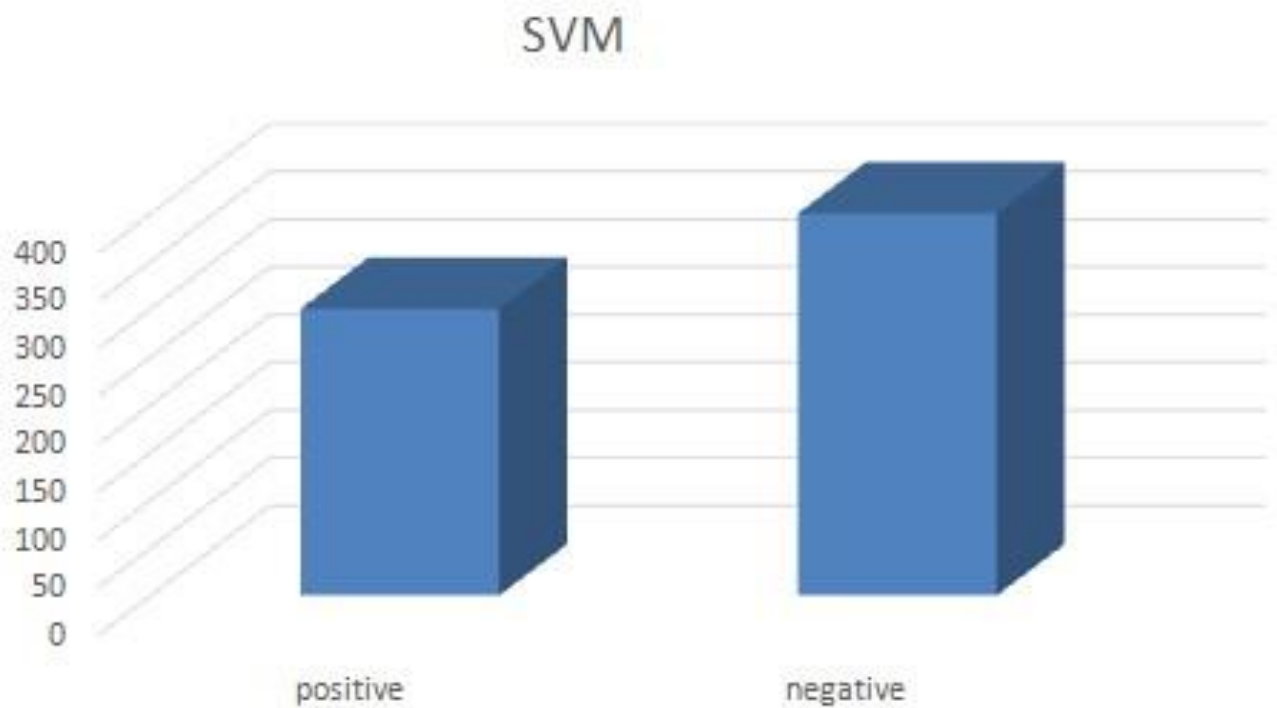
Process



Then we click **Run** to see and analyze the result as shown below:

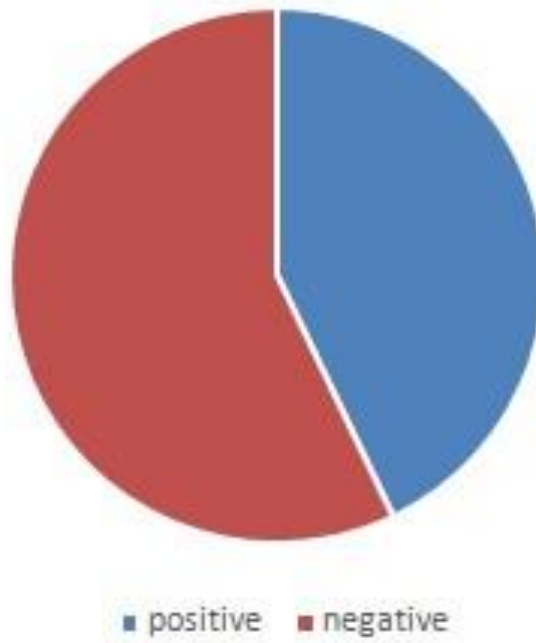
# The Result

With SVM

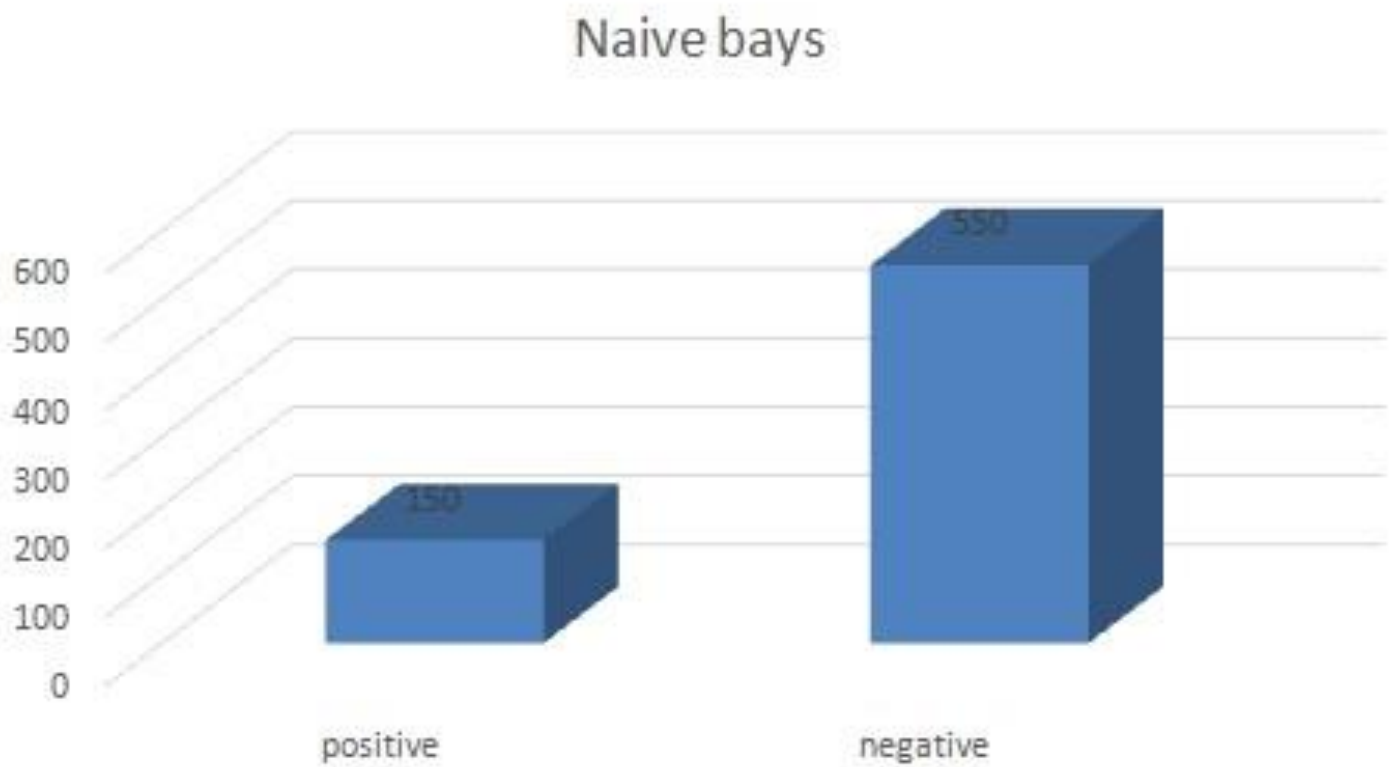




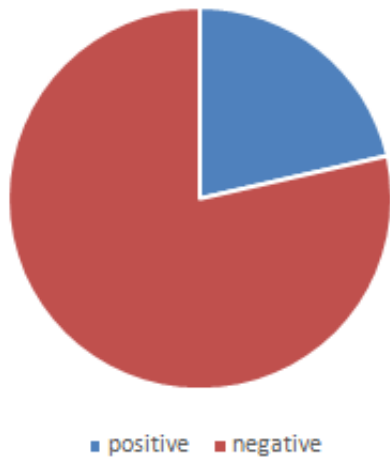
## SVM



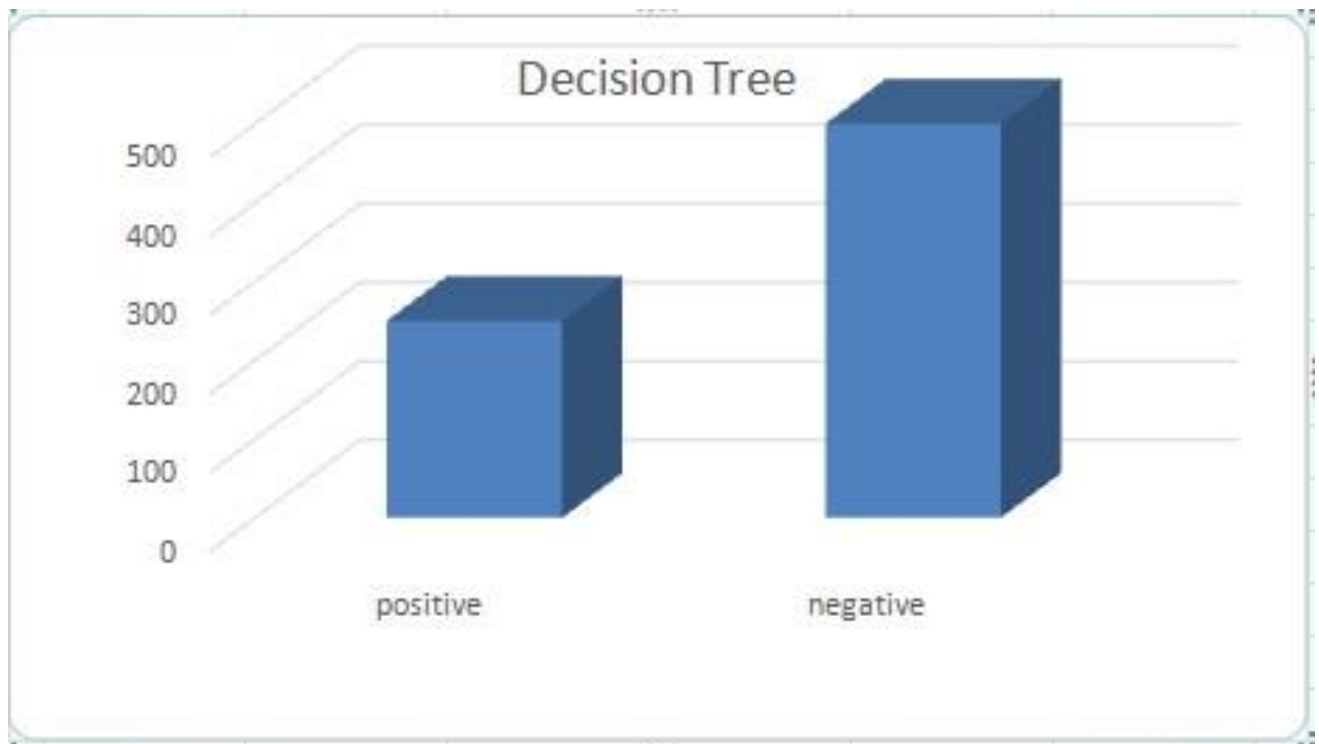
## With naïve bays



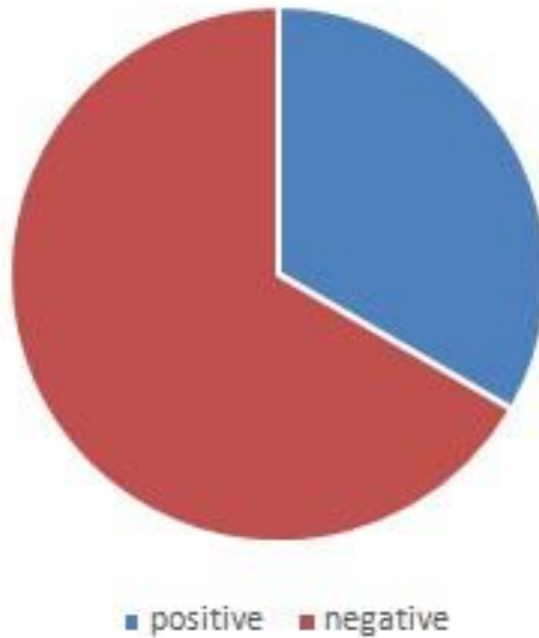
Naive bays



## Decision tree



### Decision Tree



And we notice that the three classification models are as following:

	Decision Tree	Naïve Bays	SVM
Accurecy	*****	**	***
Time Learning	***	*****	**
speed of Classification	*****	*****	*****
Recall	***	***	*****
Explanation ability	*****	***	**
binary/continuos/discret	****	***not continuos	***not discret

# For the Twitter Data


## ETL (Extract, Transform, load)

### for twitter Data

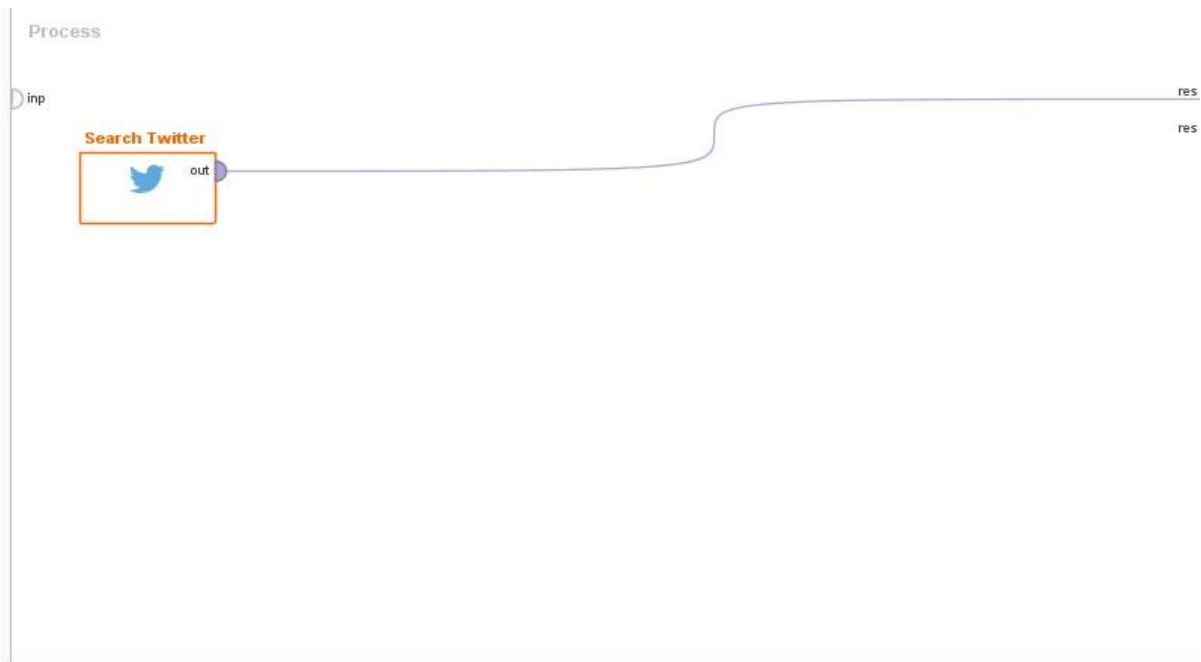
we will start with the extraction of Wal-Mart data from twitter, the tool that we use for this purpose is Rapidminer tool

### **RapidMiner tool.**

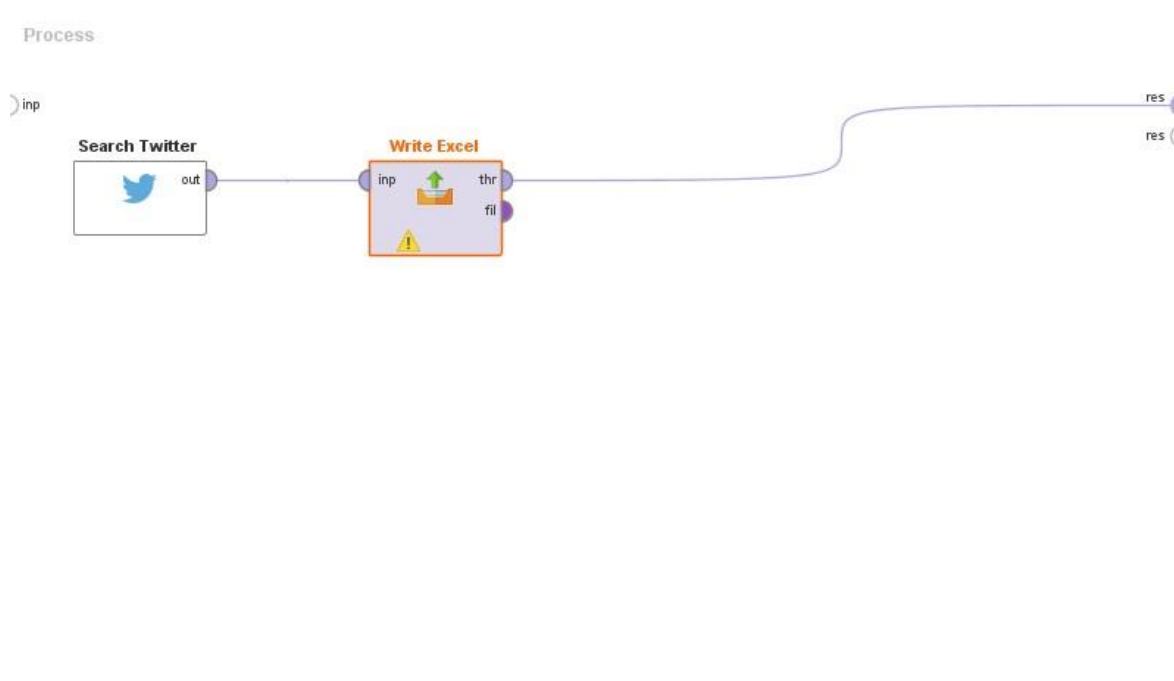
Rapid miner is a software platform for data science teams that unites data preparation, machine learning, and predictive model deployment.

We Create a new process  in RapidMiner Studio then Here we drag the “search twitter operator ” and then enter to the twitter account through the operator and then we choose the keywords that we want to search about like (Walmart , #Walmart, @Walmart , Love Walmart , #love Walmart , hate Walmart , #hate Walmart).

**As shown below:**



Before we run the process, we put it in new excel file through the process “write excel” and put all data we generate on it.

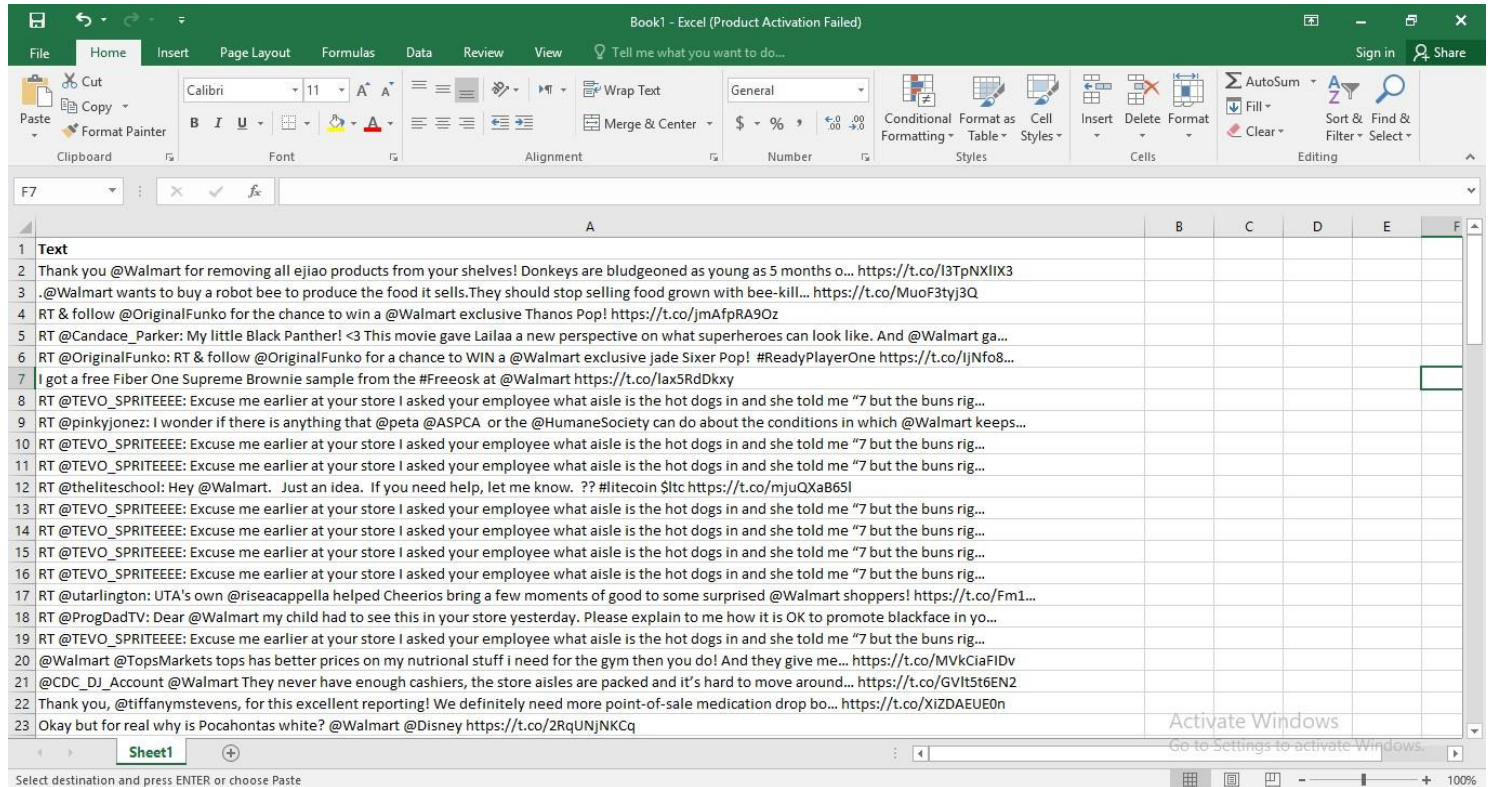


Start ----- > new connection (open Url “Request access token – login to twitter “, copy access token, test, save all changes) ----- > query ----  
 -- > limit ----- > language ----- > write excel ----- > Run

The data we generate is as shown below:

Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Text	Geo-Local	Geo-Local	Retweet-Id	Retweet-Id
2018-03-17 00:28:04	PETA	9890492.0			-1.0	en	<a href="t Thank you @Walmart for remo	159.0	974774327840264000.0		
2018-03-17 19:30:03	Friends of	19539716.0			-1.0	en	<a href="t .@Walmart wants to buy a robo	45.0	975061717020479000.0		
2018-03-13 21:02:37	Funko	1378000488.0			-1.0	en	<a href="t RT & follow @OriginalFunko fo	15781.0	973635458134626000.0		
2018-03-18 19:53:47	Agape Lov	3581952077.0			-1.0	en	<a href="t RT @Candace_Parker: My little	107.0	975430076123804000.0		
2018-03-18 19:53:45	Kyle Mato	727284014332150000.0			-1.0	en	<a href="t RT @OriginalFunko: RT & follo	5482.0	975430069656195000.0		
2018-03-18 19:53:26	CB Blog&F	457243990.0			-1.0	en	<a href="t I got a free Fiber One Supreme	.0	975429988819317000.0		
2018-03-18 19:53:22	presley	2704993662.0			-1.0	en	<a href="t RT @TEVO_SPRITEEEE: Excuse r	901.0	975429970204987000.0		
2018-03-18 19:52:54	#EdgarHO	126136711.0			-1.0	en	<a href="t RT @pinkyjonez: I wonder if th	1.0	975429852680487000.0		
2018-03-18 19:52:26	yung meri	229986822.0			-1.0	en	<a href="t RT @TEVO_SPRITEEEE: Excuse r	901.0	975429735982486000.0		
2018-03-18 19:51:27	C.	1141492476.0			-1.0	en	<a href="t RT @TEVO_SPRITEEEE: Excuse r	901.0	975429490498253000.0		
2018-03-18 19:51:25	Litecoin K	956760303257055000.0			-1.0	en	<a href="t RT @theliteschool: Hey @Waln	26.0	975429482541715000.0		
2018-03-18 19:51:05	Janay Mid	499448061.0			-1.0	en	<a href="t RT @TEVO_SPRITEEEE: Excuse r	901.0	975429396780724000.0		
2018-03-18 19:50:50	??p	168317247.0			-1.0	en	<a href="t RT @TEVO_SPRITEEEE: Excuse r	901.0	975429334377824000.0		
2018-03-18 19:50:17	Fabiansitc	806733619406389000.0			-1.0	en	<a href="t RT @TEVO_SPRITEEEE: Excuse r	901.0	975429196414566000.0		
2018-03-18 19:50:10	Melancho	349980463.0			-1.0	en	<a href="t RT @TEVO_SPRITEEEE: Excuse r	901.0	975429167885013000.0		
2018-03-18 19:49:56	Persian C	324091138.0			-1.0	en	<a href="t RT @utarlington: UTA's own @	16.0	975429105918403000.0		
2018-03-18 19:49:56	WOKEST	967819864751587000.0			-1.0	en	<a href="t RT @ProgDadTV: Dear @Walm	18.0	975429105918255000.0		
2018-03-18 19:49:35	draizy ??	3305778788.0			-1.0	en	<a href="t RT @TEVO_SPRITEEEE: Excuse r	901.0	975429018768916000.0		
2018-03-18 19:47:54	Shawn	23722811.0	Walmart	17137891.0	-1.0	en	<a href="t @Walmart @TopsMarkets tops	.0	975428597770084000.0		
2018-03-18 19:47:32	Retired N	815447261224337000.0	CDC_DJ_A	2221445064.0	-1.0	en	<a href="t @CDC_DJ_Account @Walmart	.0	975428501825343000.0		
2018-03-18 19:47:02	Laura Bow	26152056.0			-1.0	en	<a href="t Thank you, @tiffanymstevens,	.0	975428377229386000.0		
2018-03-18 19:46:59	D E S	336278605.0			-1.0	en	<a href="t Okay but for real why is Pocah	.0	975428363895607000.0		
2018-03-18 19:46:55	bjorn130	97455889.0			-1.0	en	<a href="t RT @OriginalFunko: RT & follo	5482.0	975428349945418000.0		

Then we delete all un useful columns and remain just the tweets.



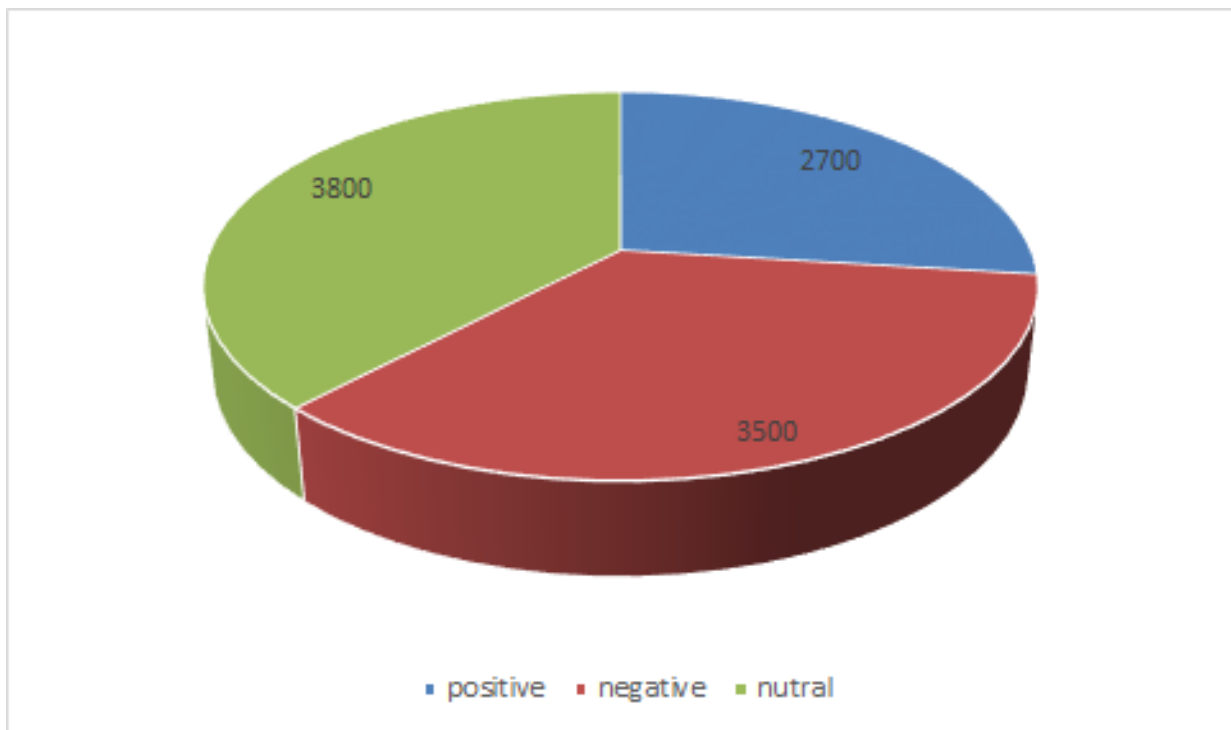
here the data that we want to work on it  
"the tweets" from twitter



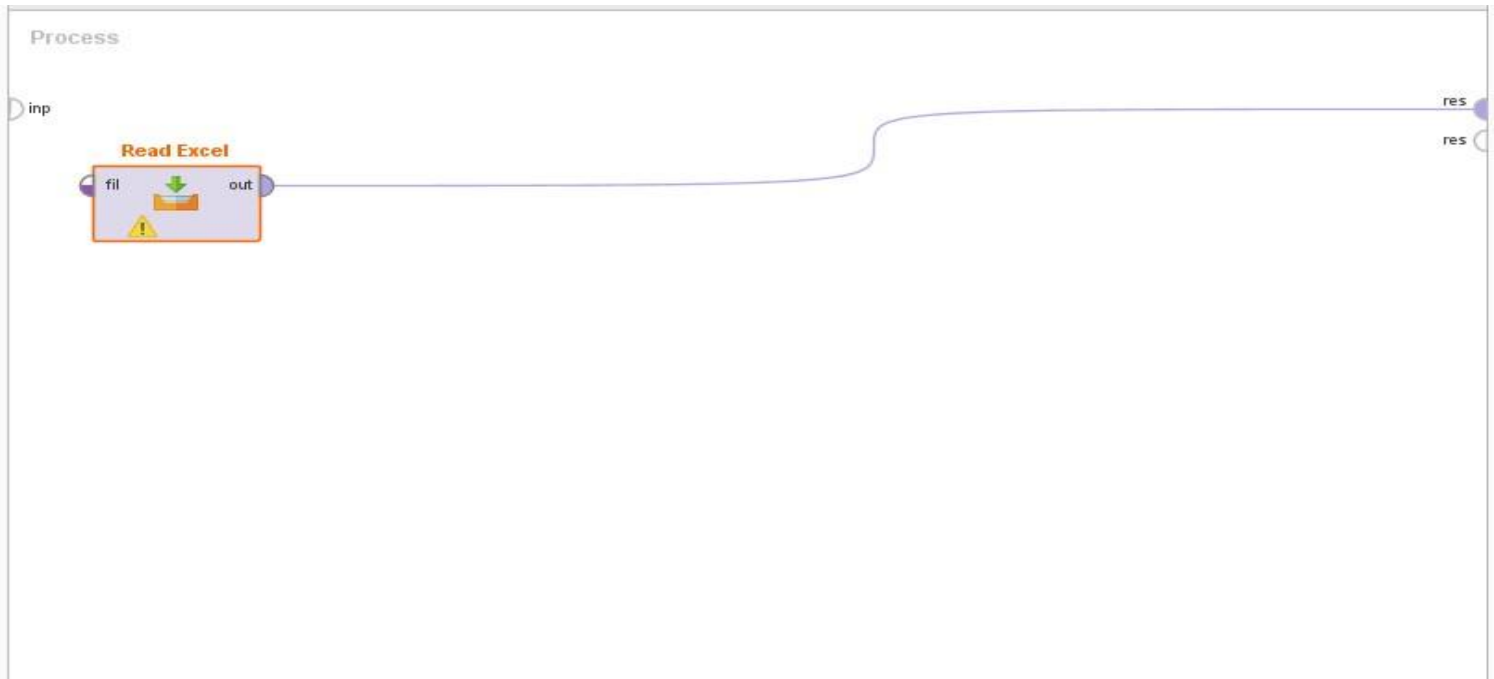
We consider to create the training data that will be used to learn the machine and make it predict the positivity and negativity off whole Walmart twitter data so **First** The data is as shown below:

A		B
1	Messeege	Result
2	Can't understand why someone would leave such a high level position to work at Walmart? One can effect rea...	negative
3	#Walmart reminding why I hate Walmart#shit store	negative
4	@_kendrapaige_ I hate Walmart in general, but that one is my least favorite.	negative
5	@AFootballLife LoL see this is another example of why I hate Walmart	negative
6	Don't forget corporate America!!! People loveeee to talk about how much they hate Walmart but won't su...	negative
7	I love going to Walmart ????	positive
8	I went to Walmart and there was a mom and her lil son and they were "competing" with "I love you!!" "I love YOU" "	positive
9	Thanks for the excellent service @walmart #grocery pickup in Estero. Kevin truly makes you feel like the most important customer of the day.	positive
10	Was working at a pretty nice walmart today and a cute little mouse ran across the floor in front of me. Gave 5 CSMs a heart attack.	positive
11	you are the best walmart go aHEAD	positive
12	morining all	neutral
13	Today I went to Walmart and there was a foundation bottle that was completely empty on the shelf	neutral
14	how are you	neutral
15	Lil Xan is at Walmart and I can't even go see him smh??	neutral
16	RT @MrGoodBeard_ : Walmart has self checkout... you can get your food for free lol	neutral

we take three classes from the beginning because we notice that there are some comments is neutral as shown previously.



the **second** step is to load the data to the rapidminer as shown below:



This operator reads an Example Set (sample data) from the specified Excel file. Our excel file contain the train data we use; we use import configuration wizard for loading data from excel file to rapid miner.

## Third:

**Set role:** This Operator is used to change the role of one or more Attributes

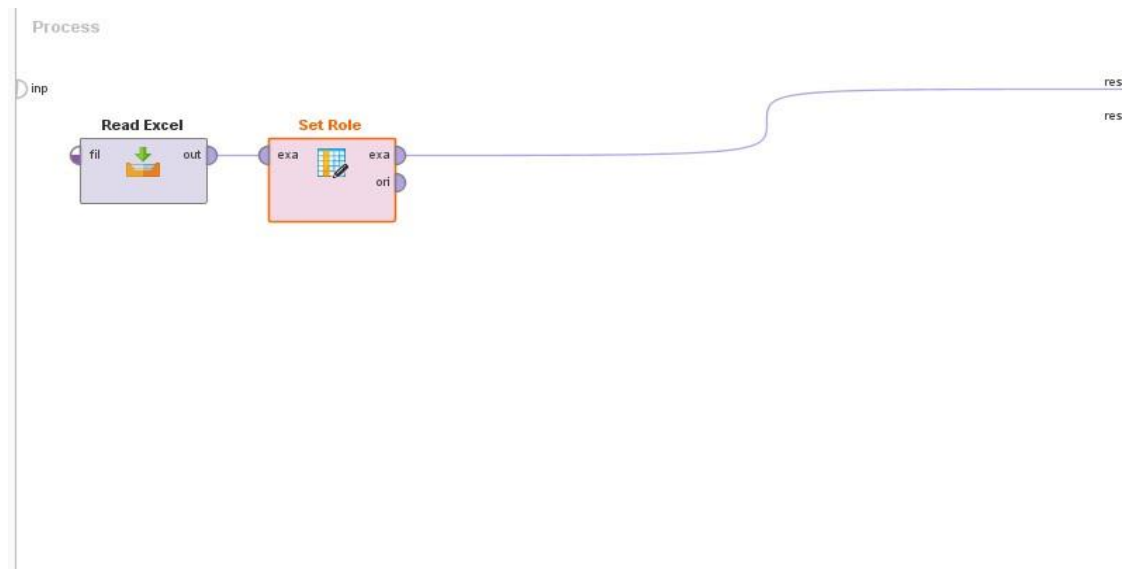
### Parameters:

**Attribute name** ->the name of the Attribute which role should be changed. The name can be selected from the dropdown menu or manual typed. (result as we name in our train file).

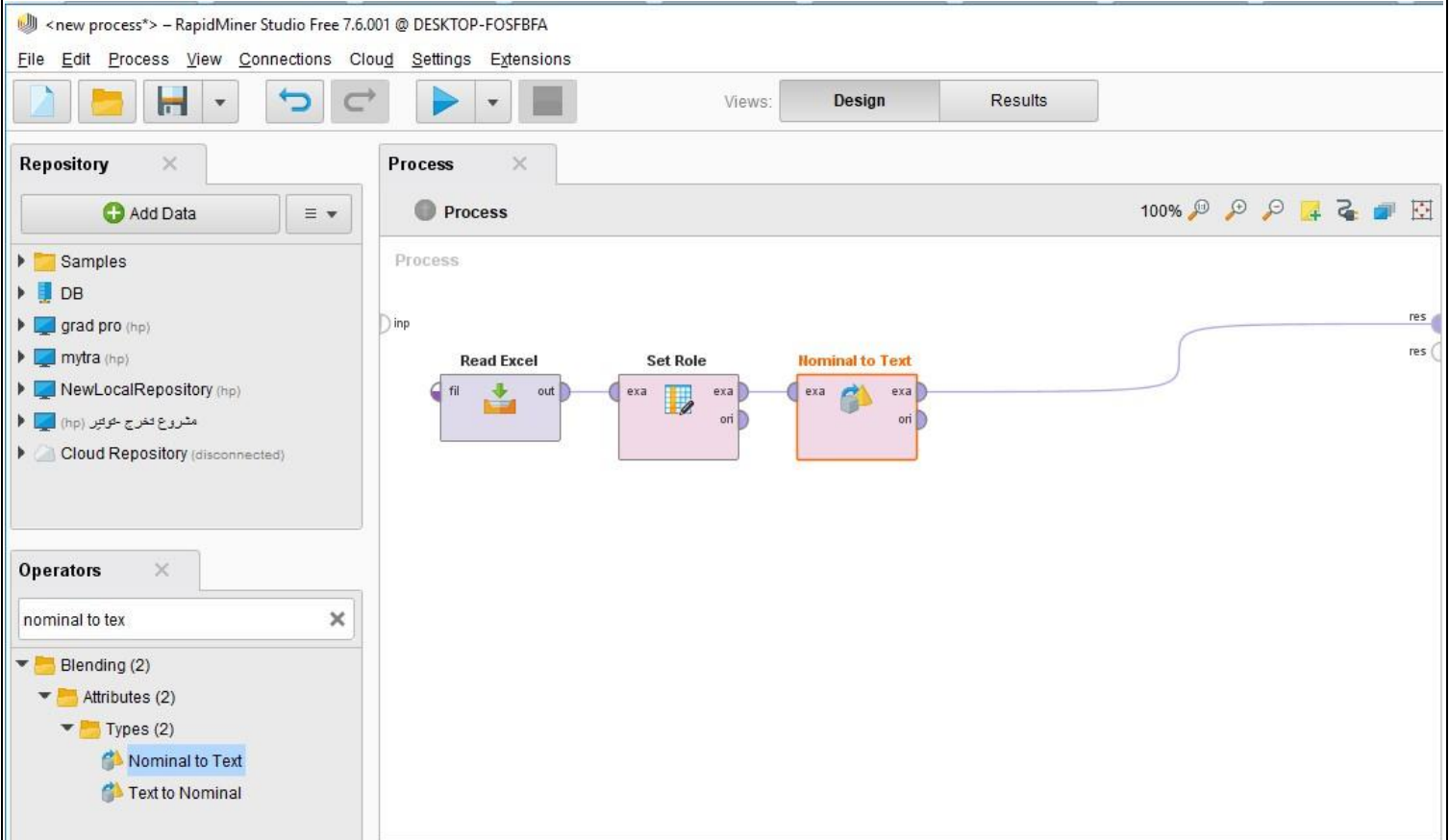
**Target role**->The target role of the selected Attribute is the new role assigned to it. Following target roles are possible:

**Label:** This is a special role. An Attribute with the label role acts as a target Attribute for learning Operators. The label is also often called 'target variable' or 'class'

.

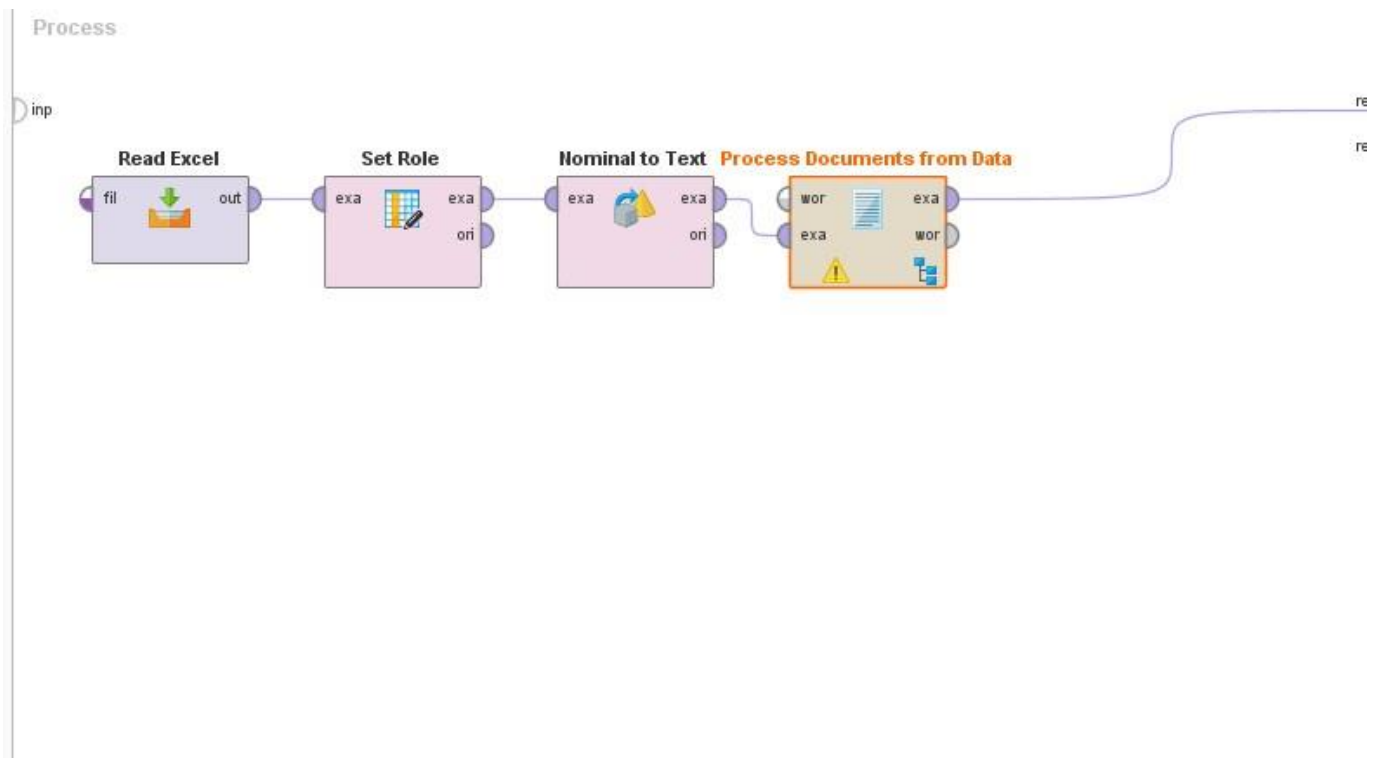


**Fourth:** we use the **Nominal to text** process (To specify which column is a text column, since Rapidminer "Process Documents..." Operators work only on text data.



## Fifth:

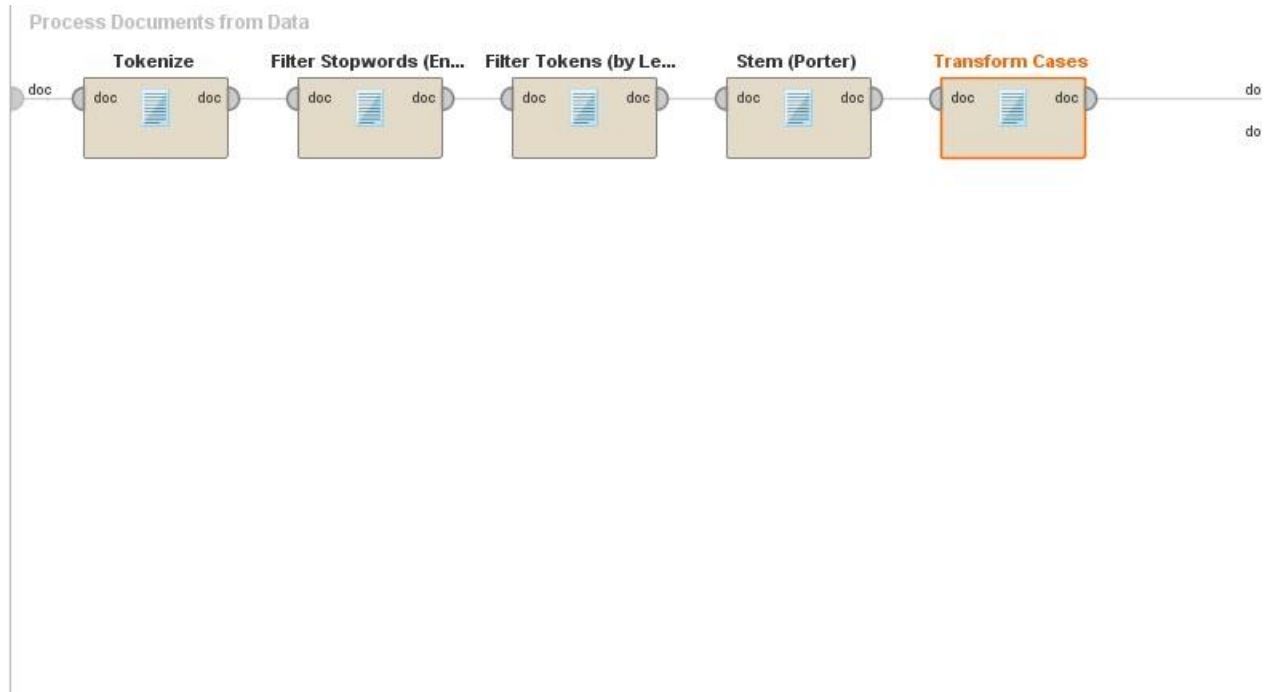
We use Process Documents from Data process:



The **Process Documents from Data** operator is used to create word vectors from text attributes, this is a nested operator that contains sub-operators inside it.

TF-IDF stands for term frequency-inverse document frequency. It is a numerical statistic which reflects how important a word is to a document.

in a collection, and it is often used as a weighting factor. And it contains sub processes like:



The **Tokenize operator** tokenizes documents, and we select in the parameters of this operator to tokenize at non letters so that each time a non-letter is found it shall denote a new token, therefore splitting a text into words (This operator splits the text of a document into a sequence of tokens)

The **Filter Stopwords (Dictionary) operator** applies a stopwords list from a file.

Stopwords are words which are filtered out prior to, or after, processing of natural language data (text).

e. For example, some of the most common stopwords for search machines are: the, is, at, which, and on.

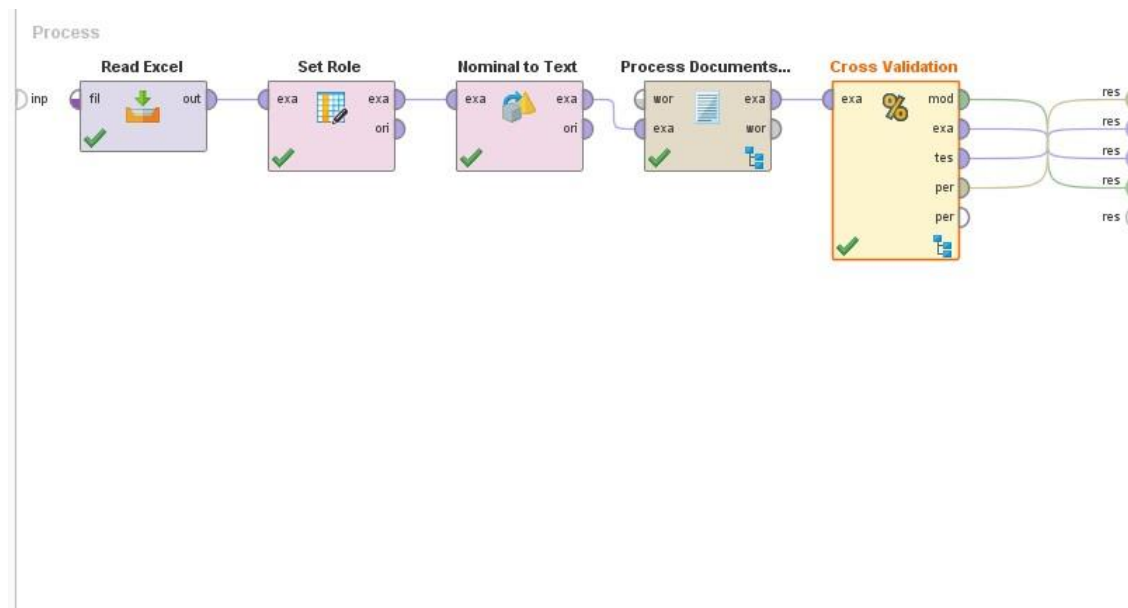
) Words that do not matter when parsing text(

The **Filter Tokens (by Length)** operator, filters tokens based on their length. In its parameters we select the min chars of a token to be 3 (thus removing single letter words), and the max chars of a token to be 20 which is safe enough to say that words consisting of 20 chars are probably gibberish.

Stemming also known as lemmatization is a technique for the reduction of words into their stems, base or root. Many words in the English language can be reduced to their base form or stem e.g. like, liking, likely, unlike belong to like.

The **Transform Cases** operator transforms the cases of all characters. In its parameters we choose to transform all characters to lower case

**Sixth:** we use the cross validation process as shown below:





**The Cross Validation Operator:** is a nested Operator.

It has two sub processes: A Training sub process and a Testing sub process. The Training sub process is used for training a model. The trained model is then applied in the Testing sub process. The performance of the model is measured during the Testing phase

$K=10$ , we choose  $k=10$  because we discover that 10 is give us a better machine Learning and a better accuracy, the model learn better What is machine learning?

Our project based on machine learning, we choose a train data and classification model and the apply of that classification model on our train data call machine learning Generally, there are 3 types of learning algorithm in our project we use Supervised Machine Learning Algorithms to make predictions we use this machine learning algorithm. Further, this algorithm searches for patterns within the value labels. That was assigned to data points. (we already know the output of our analyses (positive, negative, neutral)).

Train data = 500 row that are classify manually

Test	Train	Train	Train	Train	Train	Train	Train	Train	Train
------	-------	-------	-------	-------	-------	-------	-------	-------	-------



Max -> best validation



Max->best learning result

Number of train data sets= 500 row (comment, class)

K (number of folds) =10, number of rows (data set size) =500

$$500/10= 50$$

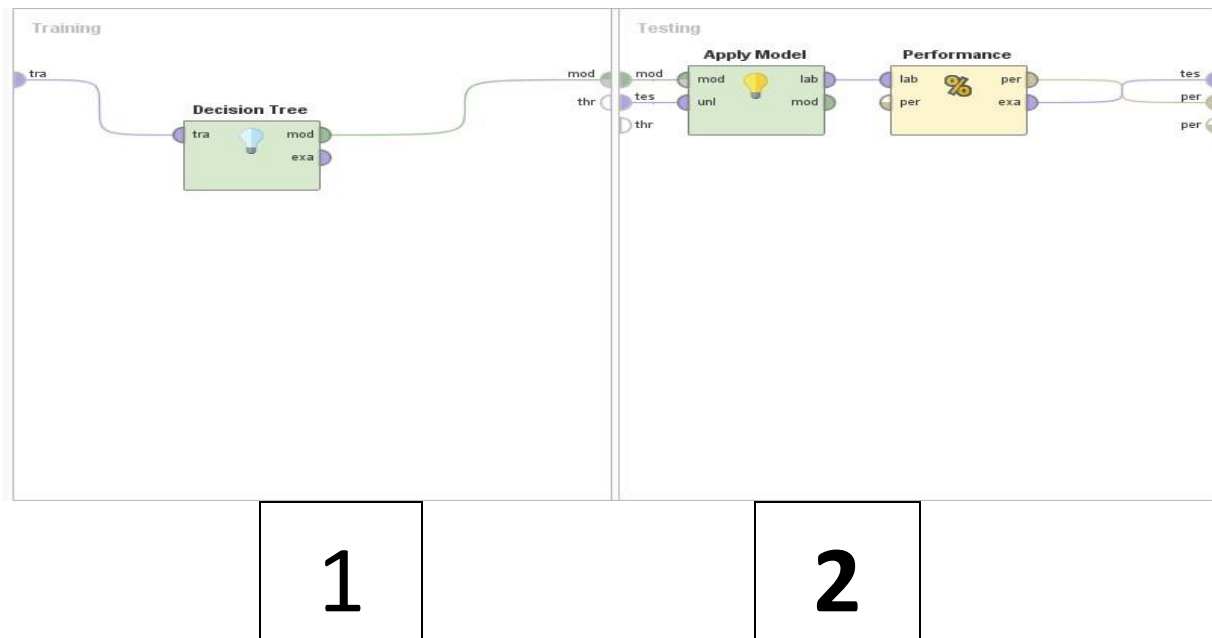
50	50	50	50	50
50	50	50	50	50

In the beginning, cross validation is a process that divides train data into parts, K is chosen as 10, and according to many experiments for more numbers we discover that number 10 is the one that gives the highest accuracy. In addition, we investigated several files of the subject and the majority advice to choose the number 10 to give better results and better learning ability. **cross validation works as follows:**

Divide the 10 groups into two parts as shown in the table above. The first section contains (9/10) groups which contains the part that the model will learn from, and the second section contains one set (1/10) which is the part that we will do a test on it and show us a certain accuracy, after that this process will make reverses groups that take 9 different groups for learning

and one group for testing process and returns to calculate the accuracy in the same way and so on (the process does the switching process until it finishes all the groups (train, test) and takes the average of All the accuracy that appeared in all operations, and shown to us as a final result rate and this is the result on which to calculate whether the accuracy is high or not.

And this process contains sub processes as shown below:




**Number 1 is the (training part)** that contains the Classification model (decision tree)

**Number 2 is the (testing part)** that contains: -

1- The apply model (This Operator applies a model on an Example Set (train set)).

2- The performance operator (This operator is used for performance evaluation. It delivers a list of performance criteria values. These performance criteria are automatically determined in order to fit the learning task type.)

After We click on RUN (  ) we can notice the program run by see the cross validation operator

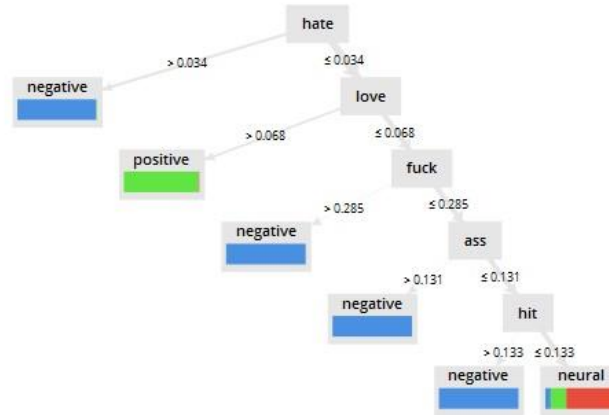
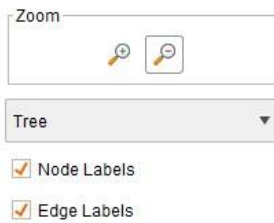
# THE RESULT

accuracy: 83.06% +/- 5.56% (mikro: 83.08%)

	true negative	true positive	true neural	class precision
pred. negative	126	1	1	98.44%
pred. positive	0	103	2	98.10%
pred. neural	24	49	149	67.12%
class recall	84.00%	67.32%	98.03%	

Row No.	Result	prediction(R...	confidence(...	confidence(...	confidence(...	aaroncart	abandon	abbygailram...	absi
1	negative	negative	0.991	0	0.009	0	0	0	0
2	negative	negative	0.991	0	0.009	0	0	0	0
3	negative	neural	0.077	0.699	0.224	0	0	0	0
4	negative	neural	0.077	0.699	0.224	0	0	0	0
5	negative	negative	0.991	0	0.009	0	0	0	0
6	negative	negative	0.991	0	0.009	0	0	0	0
7	negative	negative	0.991	0	0.009	0	0	0	0
8	negative	negative	0.991	0	0.009	0	0	0	0
9	negative	negative	0.991	0	0.009	0	0	0	0
10	negative	negative	0.991	0	0.009	0	0	0	0
11	negative	negative	0.991	0	0.009	0	0	0	0
12	negative	negative	0.991	0	0.009	0	0	0	0
13	negative	negative	0.991	0	0.009	0	0	0	0
14	negative	negative	0.991	0	0.009	0	0	0	0
15	negative	negative	0.991	0	0.009	0	0	0	0

# Decision tree



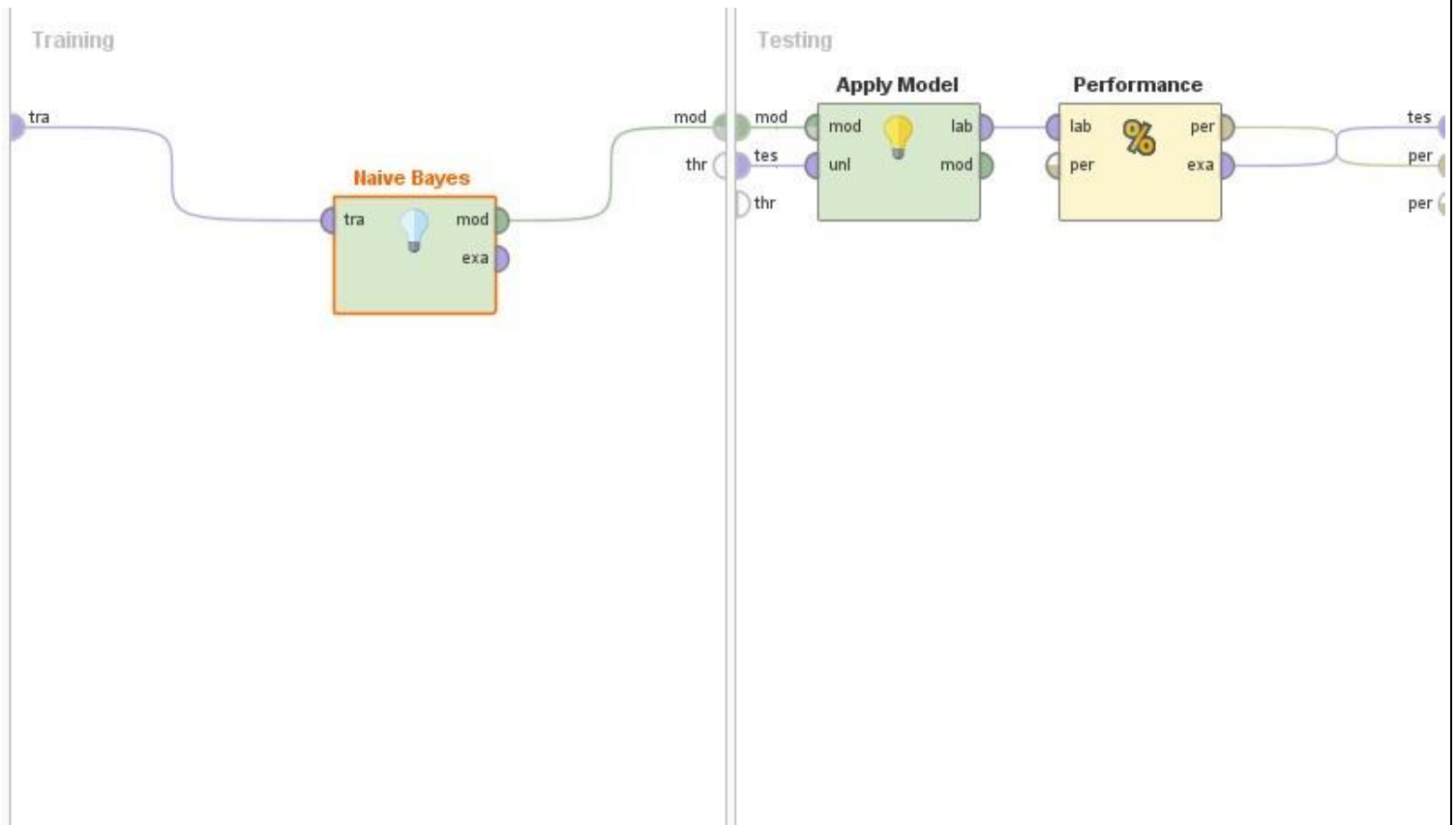
Now we want to try a different classification model to see if the accuracy change (increase, decrease) to decide which classifier is better to use.

**First** we try by using decision tree

## Description

```
hate > 0.034: negative {negative=126, positive=1, neural=0}
hate ≤ 0.034
| love > 0.068: positive {negative=0, positive=103, neural=1} |
love ≤ 0.068
| | fuck > 0.285: negative {negative=3, positive=0, neural=0}
| | fuck ≤ 0.285
| | | ass > 0.131: negative {negative=2, positive=0, neural=0} |
| | | ass ≤ 0.131
| | | | hit > 0.133: negative {negative=2, positive=0, neural=0}
| | | | hit ≤ 0.133: neural {negative=17, positive=49, neural=151}
```

Second we want to choose **naïve bays** classifier





The **confusion matrix** is as shown below:

☒ Table View ☐ Plot View

accuracy: 53.17% +/- 6.02% (mikro: 53.19%)

	true negative	true positive	true neural	class precision
pred. negative	99	50	46	50.77%
pred. positive	25	76	39	54.29%
pred. neural	26	27	67	55.83%
class recall	66.00%	49.67%	44.08%	

## Description

Performance Vector:

Accuracy: 53.17% +/- 6.02% (mikro: 53.19%) Confusion Matrix:

True:    negative positive neural

Negative:        99        50        46

Positive: 25 76        39 neural:        26  
             27        67

kappa: 0.298 +/- 0.090 (mikro: 0.298) Confusion Matrix:

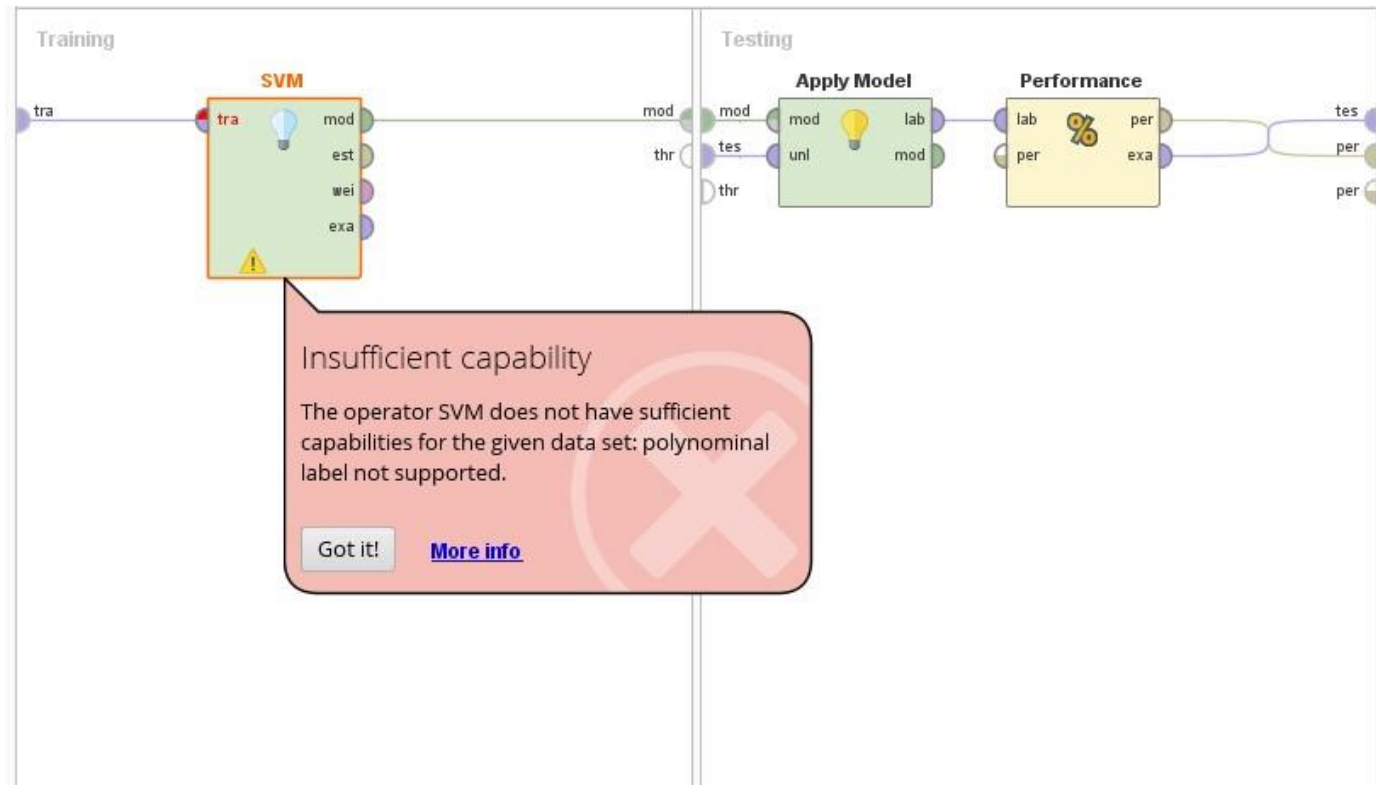
True:    negative positive neural

negative: 99        50        46

positive: 25        76        39

neural: 26        27        67

Third we use **SVM** (support vector machine)



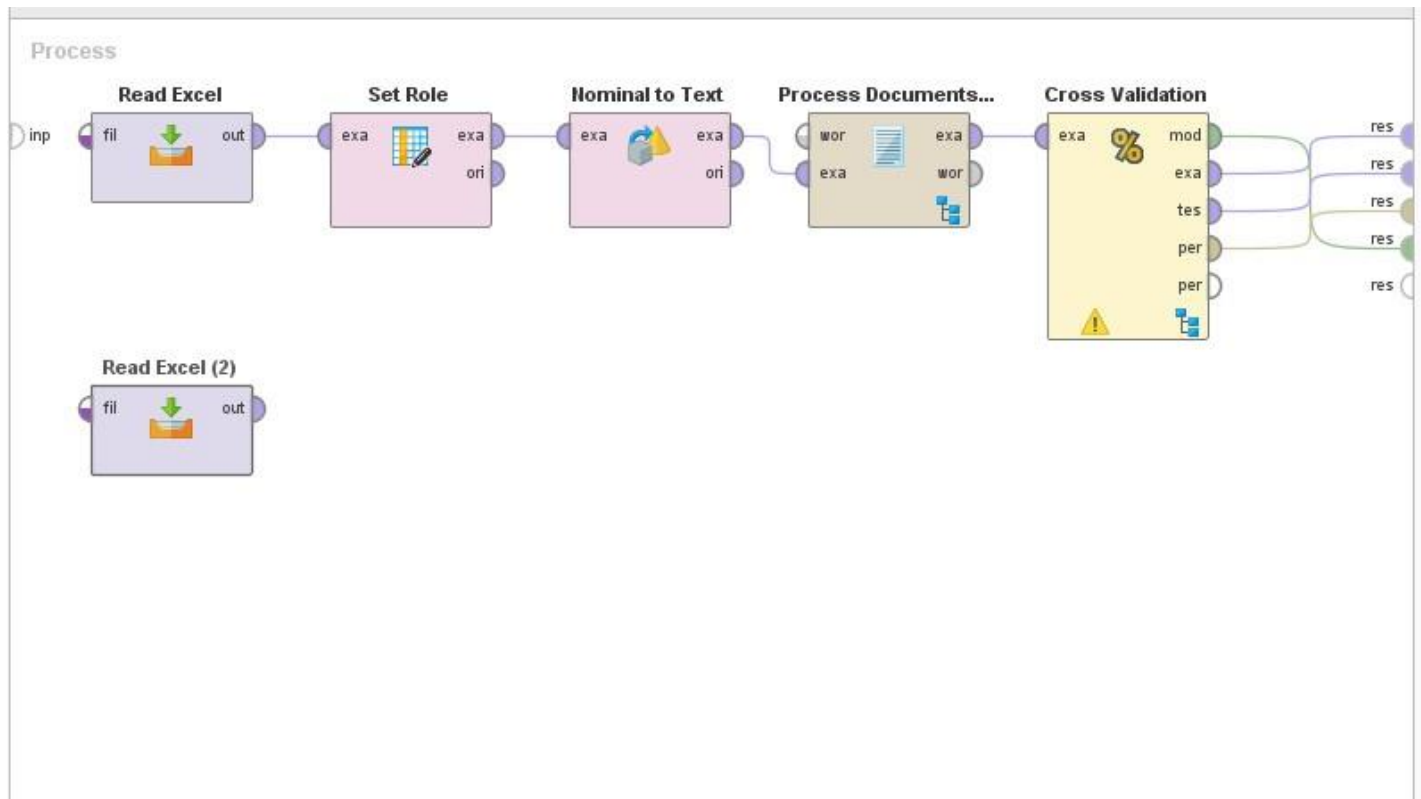
As we see the error, SVM does not support polynomial label (positive, negative, neutral).

Model	Accuracy
Decision Tree	83.06%
Naïve Bayes	53.17%
SVM	Does not support

Then we will let the machine execute what it learns from the training data we give for all data generated from the twitter about Walmart

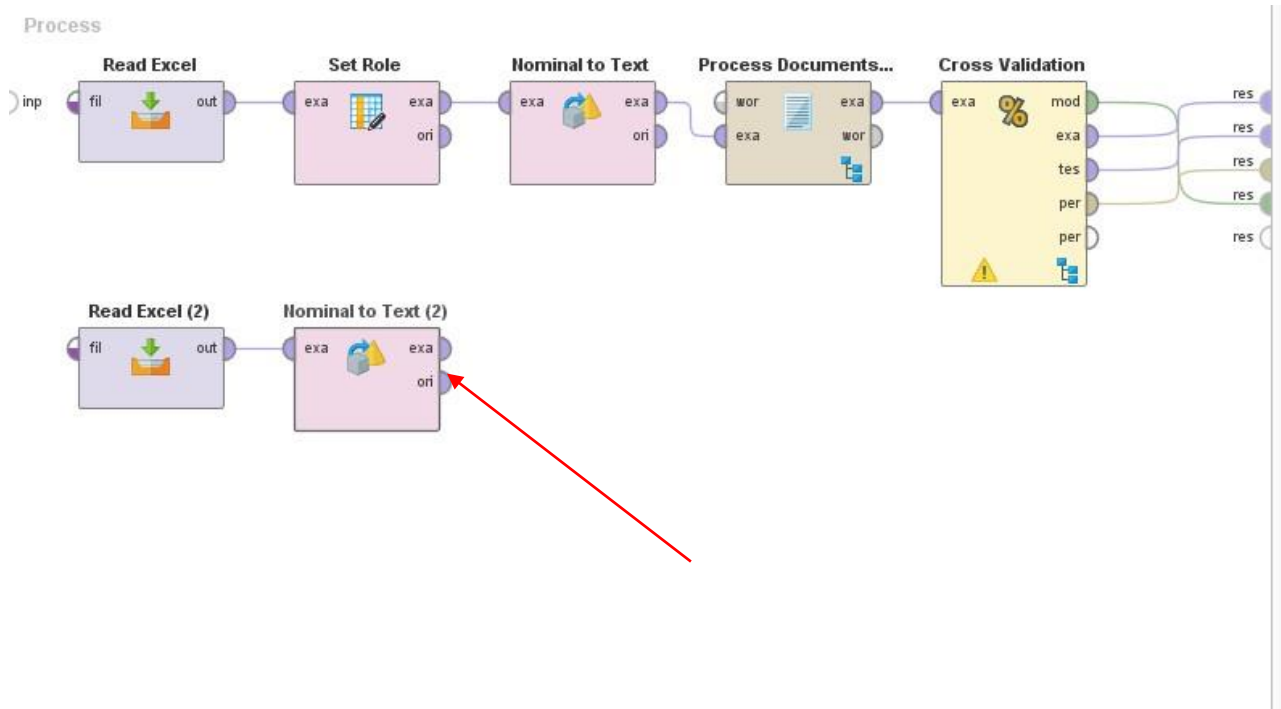
## First:

We load all twitter Walmart data into the Rapidminer using the read Excel process like shown below:



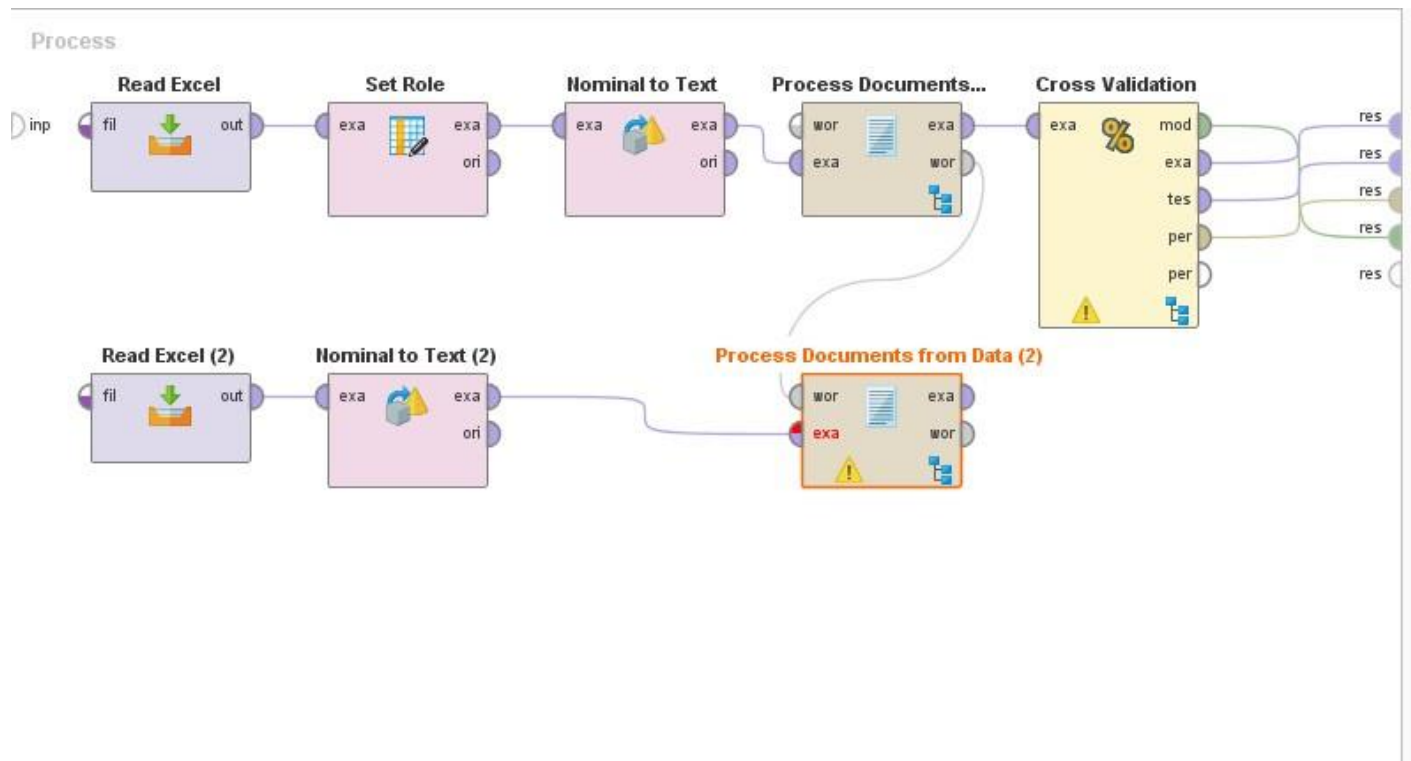
## Second:

we use the **Nominal to text** process (To specify which column is a text column, since Rapidminer "Process Documents..." Operators work only on text data **as shown below**:



## Third:

We use Process Documents from Data process that we used previously and we use the same sub process shown and described previously and they look like this:

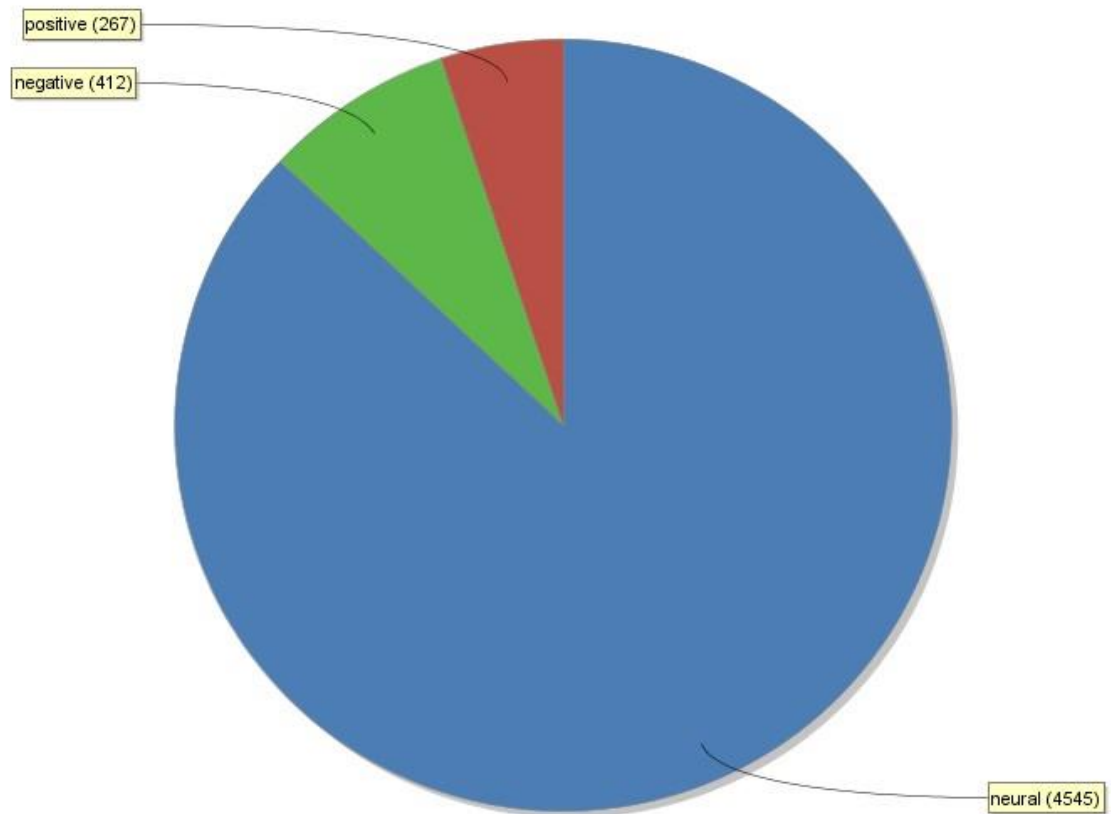
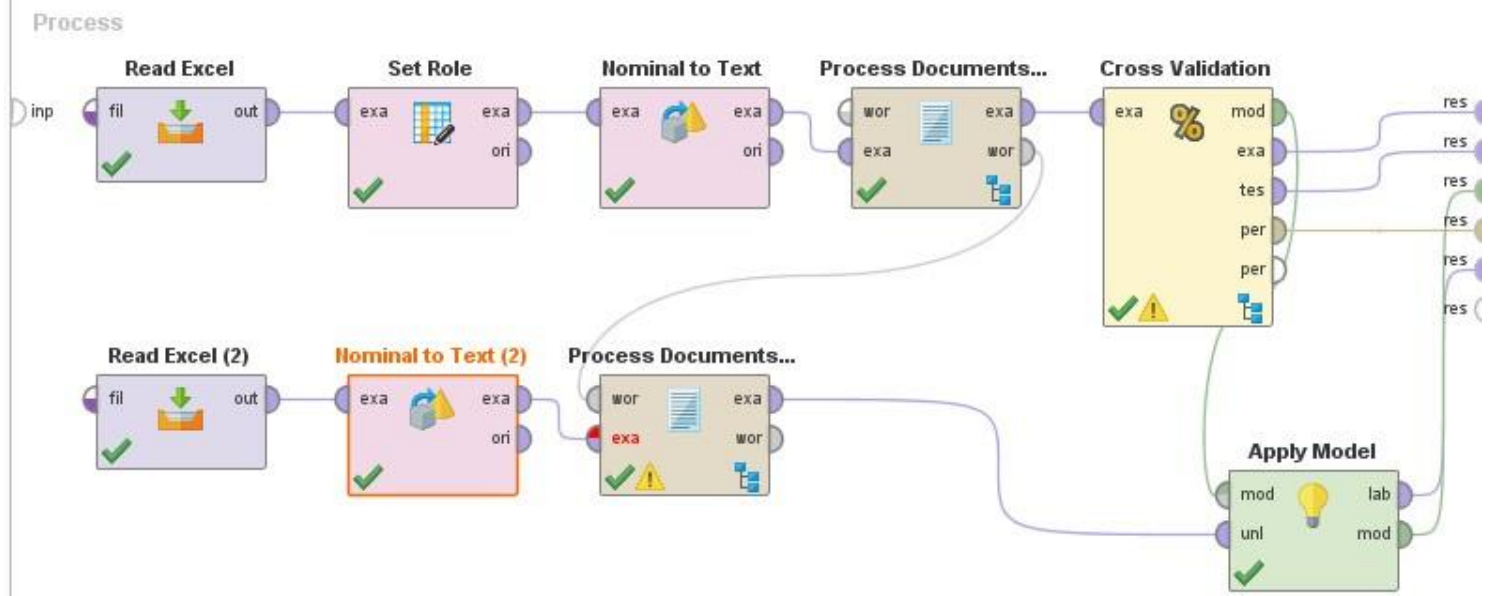


And the sub processes are:



## Forth:

We choose apply model operator to execute all above processes as shown below:



## Problems that faces us in the project:

1. Tool not fully support some languages such as Arabic.
2. Tool does not support the huge number of data -
3. the existence of most comments in the colloquial language which is not analyzed by neither humans or tools.
4. tool must be paid to give better results in pulling data.
5. lack of sufficient resources to learn through.
6. The next step to work we were trying to find it with difficulty because the idea is new.
7. The difficulty of finding tool that performs the process of pulling the data fully and well.

## Solutions & recommendations:

After reviewing and studying results, recommendations had to be made to solve negative problems related to the opinions of people on certain topics of Walmart:

1. they should pay more attention to the website and e-procurement because of many complaints related to Online Service.
2. there are also some customers who have complained about the bad treatment of the employees.



- 3. disturbing customers by the fact that the branches of Walmart have full of work pressure and crisis forcing customers to wait.
- 4. some complain that customers accuse Walmart of not paying attention of negative comments on social media and sometimes deleting them instead of dealing with them.
- 5. Online Chat does not work on their site as customers are not answered.
- 6. We also recommend companies to do the sentiment analysis process in order to know their customers' opinions.
- 7. We also have the ability to analyze any company.

# Thank you

