

**Najah National University**

**Faculty of Engineering &**

**Information Technology**

**Management Information**

**System Department**

**Graduation Project of**

**Documents Classification System**

**Classify BBC & Reuters News Using RapidMiner Tool / Text Mining**

**Development Team:**

**Razan Abdulhaq**

**Shuruoq Adawi**

**Rasha Deriyieh**

**Supervised By:**

**Dr. Kamal Irshaid**

**Dr. Maher Abu Baker**

**2019-2020**

**Table of Contents:**

1. INTRODUCTION–––––––––––––––––––––––––––––––––––––––––––2

1.1 DATA MINING.–––––––––––––––––––––––––––––––––––––––2

1.2 TEXT MINING.––––––––––––––––––––––––––––––––––––––– 3

1.3 TEXT MINING VS DATA MINING.–––––––––––––––––––––– 3

1. DOCUMENTS CLASSIFICATION–––––––––––––––––––––––––––– 4

2.1 MACHINE LEARNNG SOFTWARE (ML):––––––––––––––––6

2.1.1 MACHINE LEARNING METHODS:––––––––––––––––– 6

2.2 WHY DOCUMENTS CLASSIFICATION ARE USEFUL?–––– 9

2.2.1 The Arabic Survey.––––––––––––––––––––––––––––––– 9

2.2.2 The English Survey.–––––––––––––––––––––––––––– 10

2.2.3 The Surveys Results.––––––––––––––––––––––––––––– 11

1. TOOLS WERE USED IN THE PROJECT.–––––––––––––––––––– 11

3.1 RAPIDMINER TOOL.–––––––––––––––––––––––––––––––– 12

3.2 KUTOOLS FOR EXCEL.––––––––––––––––––––––––– –––– 13

3.3 CORVID EXPERT SYSTEM DEVLOPMENT TOOL –––––– 14

1. TEXT MINING PROCESS :––––––––––––––––––––––––––––––– 14

4.1 DATA AQUESITION.––––––––––––––––––––––––––––––––– 15

4.2 TEXT PREPROCCESSING.–––––––––––––––––––––––– ––– 20

4.3 TEXT MODELING.––––––––––––––––––––––––––––––– ––– 24

4.4 TESRING AND EVALUEATION.–––––––––––––––––––– –– 28

4.5 APPLICATION–––––––––––––––––––––––––––––––––––––– 39

5. PROBLEM AND RECOMMENDATION –––––––––––––––––––––– 41

5.1 PROBLEMS PRESENT DURING PROJECT–––––––––––––– 41

5.2 RECOMMENDATION ––––––––––––––––––––––––––––––– 41

**1. INTRODUCTION:**

The exponential growth of the Internet has led to a great deal of Interest especially by Companies in developing useful and efficient Tools or Software’s to assist employees for doing their job and users for searching the Web.

Data Mining, Machine Learning, Artificial Intelligence, Natural Language Processing (NLP), Text Mining and many other terms were appearing according to this growth.

**1.1 DATA MINING:**

Is the process of analyzing hidden Patterns of Data according to different perspectives for Categorization into useful Information which is collected and assembled in common areas, such as Data Warehouses and Structured Database; for an Efficient Analysis, Data Mining Algorithms, Decision Making and other Information requirements in order to ultimately cut Costs and increase Revenue.

**Data mining** is also known as a Data Discovery and Knowledge Discovery and it has a different Areas such as Data Mining such as Artificial Intelligence, Statistics, Database, Machine Learning, Pattern Recognition and Text Mining.

2

**1.2 TEXT MINING:**

Is the process of Exploring and Analyzing large amounts of Data aided by Software that can identify Concepts, Patterns, Topics, Keywords and other attributes in the Data. It's also known as Text Analytics.

It’s designed to help the Business to find out valuable knowledge from a text **Based on Content .**These contents can be collected in a form of Word Documents, Emails, CSV Files, XML Files, Social Media and more.

**1.3 TEXT MINING VS DATA MINING:**

Text Mining and Data Mining are often used interchangeably to describe how Information or Data is processed.

**The similarity between these two concepts are:**

**First,** both of them are seek for discover a **novel and useful patterns from Data** Such as Classification, Clustering, Association and Prediction.

**Secondly,** both of them are a **Semi-Automated Processes** which means a procedure that has yet to be completely automated; it still requires a little of real Human Interaction.

**The only Difference between them is in the Nature of the Data that can be a Structured versus Unstructured Data:**

* **Structured data in Data Mining:** which is a highly organized and easily understood by Machine language.

3

* Those working within Relational databases or Data Warehouse can input, search, and manipulate structured Data relatively quickly which is the most attractive feature of structured data .There is many example about Structured Data such as Names, Dates , Addresses, Credit Card Numbers, Stock Information and more .
* **Unstructured Data in Text Mining**: Its most often categorized as qualitative data which mean that it cannot be processed and analyzed using conventional Tools and Methods and also it hasn’t any pre-defined model, meaning it cannot be organized in relational databases that make it more difficult to deconstruct. There is many example of it include text, video, audio, mobile activity, social media, Word documents, PDF files, text excerpts, XML files, and so on
* **TEXT MINING –FIRST, THEN IMPOSE STRUCTURE TO THE DATA, FINALLY MINE THE STRUCTURED DATA.**
* **MANY APPLICATION AREAS ARE APRES FOR TEXT MINING SUCH AS INFORMATION EXTRACTION, TOPIC LAPELING, SUMMARAIZATION, TEXT CLASSIFICATION, QUESTION ANSWERING AND MORE.**

**2. DOCUMENTS CLASSIFICATION:**

Today the majority of available Business Data is Unstructured Information especially in the form of text which became an extremely rich source of Information, but extracting insights from it can be hard and time-consuming due to its unstructured nature. So Businesses are turning to Text Classification for Structuring Text in a fast and cost-efficient way to enhance Decision-Making and Automate processes.

4

**Text Classification**: which is also known as a text categorization or text tagging) is the task of assigning a set of predefined categories to free-text , Text classifiers can be used to organize, structure, and categorize pretty much anything.

**There are many broad Applications of Text Classification such as** Sentiment Analysis, Topic labeling, Document Classification and Spam Detection.

So according to these Benefits and the Important of Text Classification the idea about our project is began so we decided to focus in one of Text Classification Application – Document Classification - and build **a Document Classification System.**

**Document Classification System or Document Categorization System:**

a task of assigning a [document](https://en.wikipedia.org/wiki/Document) which may be (Audio , Text , Images ,Music) to one or more [classes](https://en.wikipedia.org/wiki/Class_(philosophy)) or [categories](https://en.wikipedia.org/wiki/Categorization) making it easier to manage and sort. It may be done "manually" (“intellectually") or [algorithmically](https://en.wikipedia.org/wiki/Algorithmically) using different text mining tool.

So For a large companies **such as** a News Sites, Municipalities , Police Centers, Universities Websites and Libraries or anyone whom deal with a lot of contents and documents that need to be classified our system will be very useful to reduce the time , efforts and make their work done more effectively and efficiency .

5

In our project we focused on **text** type of documents and took Reuters and British Broadcasting Corporation (BBC) News Sites as a Case studies to build our System.

**2.1 MACHINE LEARNNG SOFTWARE (ML):**

**Machine learning** is an Application of Artificial Intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.

* The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.

**2.1.1 MACHINE LEARNING METHODS:**

Machine learning algorithms are often categorized as Supervised or Unsupervised:

6

* **Supervised machine Learning :** it’s an algorithm that can apply what has been learned in the past to new data using labeled examples to predict future events.
* Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
* Can be further grouped into **Classification and Regression** problems.
* **In contrast, unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled.
* Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
* Can be further grouped into **Clustering and Association** problems.

7

So in our project We deepened on \***Supervised Machine Learning Algorithm\*** by collecting a large amount of Documents from the BBC and Reuters News sites which have already been correctly Classified to one for one of a five Categories which is Sports, Technology, Politics, Business and Entertainments **(**which is called a **Train Data)**

We also have a documents that haven't been classified**(Test Data)**.So our goal was to classify these documents into one of the categories that we mentioned previously by using one of text mining classification Tool which called **RapidMiner Tool.**

* Machine learning enables Analysis of massive quantities of Data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly.
* Combining machine learning with Human Abilities can make it even more effective in processing large volumes of information.

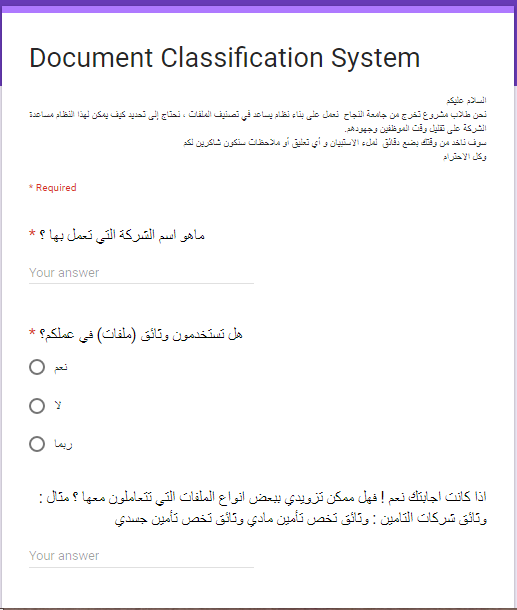
**2.2 WHY DOCUMENTS CLASSIFICATION ARE USEFUL?**

Before we started our project, we wanted first to validate if it will be useful for the organizations or it’s just like any other project related to these fields.

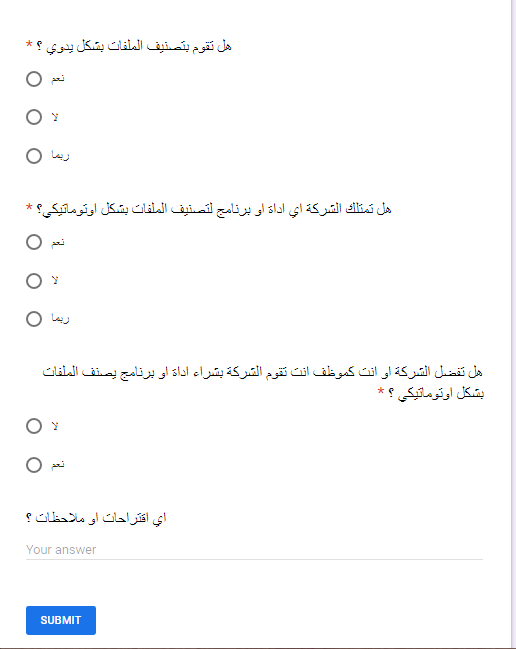
8

So we published a Survey on the Internet and we took a West Bank as a Sample and some surveys were filled manually for the companies who didn’t fill our survey such as Policies Centers and Ministries.

Then we sent a surveys to some international company such as Amazon, BBC and Reuters News agencies Company because we wanted to take a news agencies as a Case study for our project.

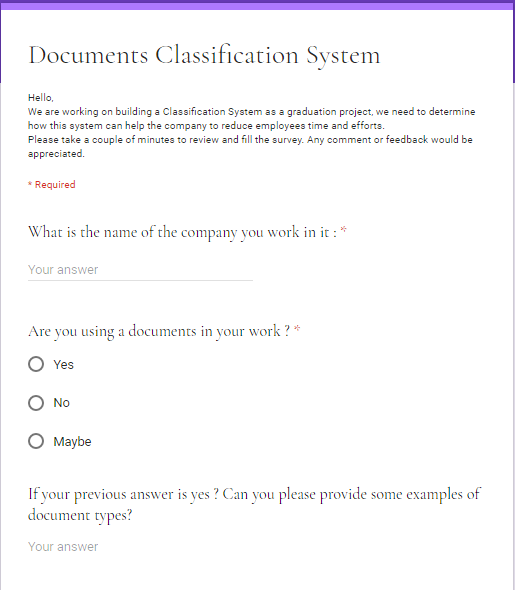
**2.2.1 The Arabic Survey**:

9

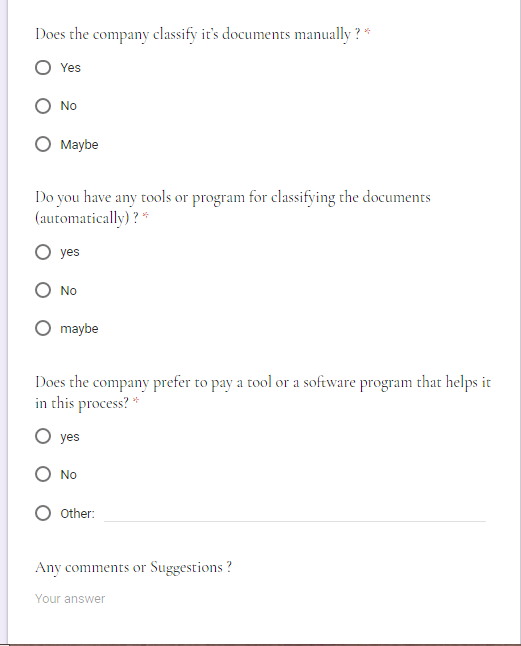
****

9

**2.2.2 The English Survey:**

****

11

****

10

**2.2.3 The Surveys Results:**

The Surveys results was 94% of Companies deal with Documents from a different types and 80% don’t have a system to classify their Documents and 85% preferred to have a system or a tool to help them in doing that .

**So the Documents Classification System is useful because?**

* Its help to Classify Document Automatically and quickly.
* Reduce the Efforts and Time of Employee.
* Ease of handling large amounts of Data.
* Help an organizations to do it works more effectively.

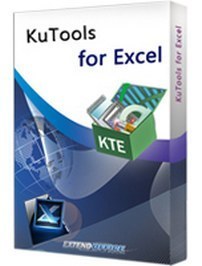
**3. TOOLS WERE USED IN THE PROJECT:**

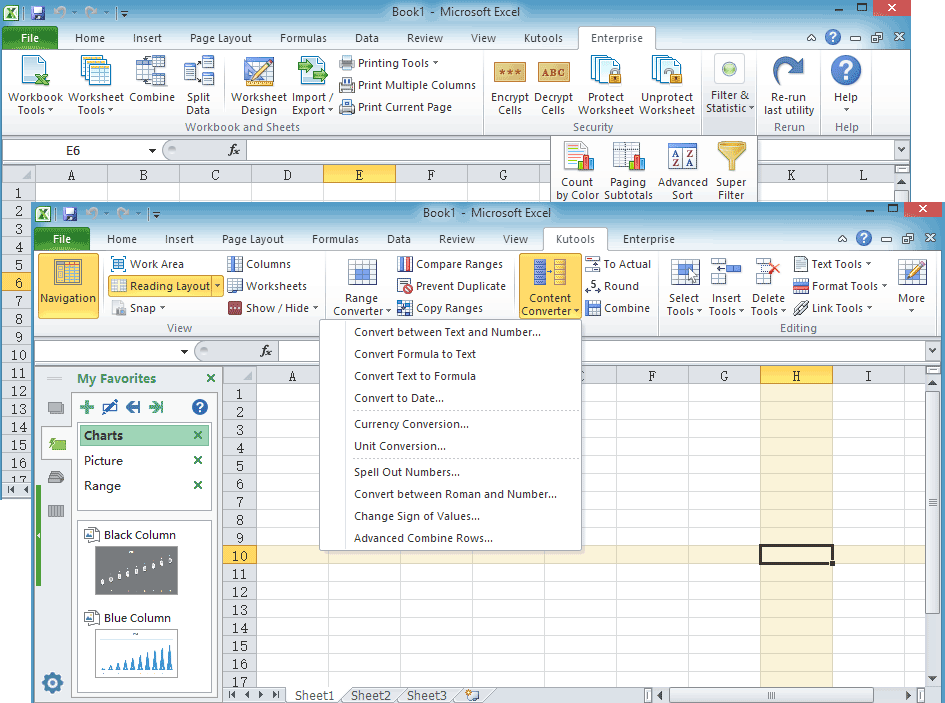
**3.1. RAPIDMINERTOOL:**

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, and Text Mining and Predictive Analytics**.** And it’s also a free open source and commercial product for Text Mining, a world-leading open-source system for Data Mining. It is available as a stand-alone Application for Data Analysis and as a Data Mining engine for the integration into own products.

11

RapidMiner is now RapidMiner Studio and Rapid Analytics is now called RapidMiner Server. It’s 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

**3.2. KUTOOLS FOR EXCEL:**

****

12

14

It’s a powerful add-in that frees you from performing time-consuming operations in Excel, such as combine sheets quickly, merge cells without losing data, paste to only visible cells, count cells by color and so on. 300+ powerful features / functions for Excel 2019, 2016, 2013, 2010, 2007 or Office 365!

**3.3- CORVIDEXPERT SYSTEM DEVLOPMENT TOOL**

**Corvid System** is a very powerful environment for developing knowledge Automation Systems, It allows the logical Rules and Procedural steps used to make a Decision to be converted to a "Rule" representation that can be delivered on-line. It is not difficult to learn to build rules with Exsys Corvid.

Corvid uses an English and Algebra syntax to build structured If/Then. Rules are easy to read and understand. No Programming Knowledge is required.

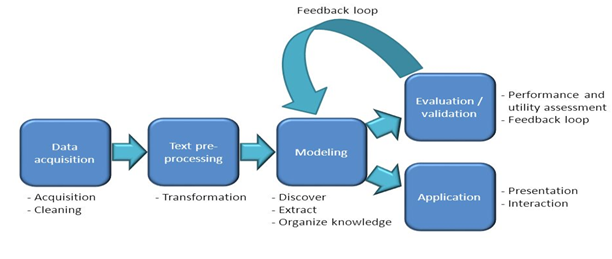
Corvid is the result of over 28 years working with Businesses and organizations to create a Tool based on what developers need to build Real world Systems. It takes a very pragmatic approach to System development make it easy for developers to get systems built and fielded as quickly as possible. Using the online tutorials, most users are able to start building small systems in a few hours. Despite its easy-to-lean nature, Corvid has been used to build highly complex systems.

13

**Corvid solves the 3 main Issues in Expert System Development:**

* Fully capturing the decision-making logic and process of the domain expert
* Wrapping the system in a user interface with the desired look-and-feel
* Integrating with other IT resources

**4- Text Mining Process:**



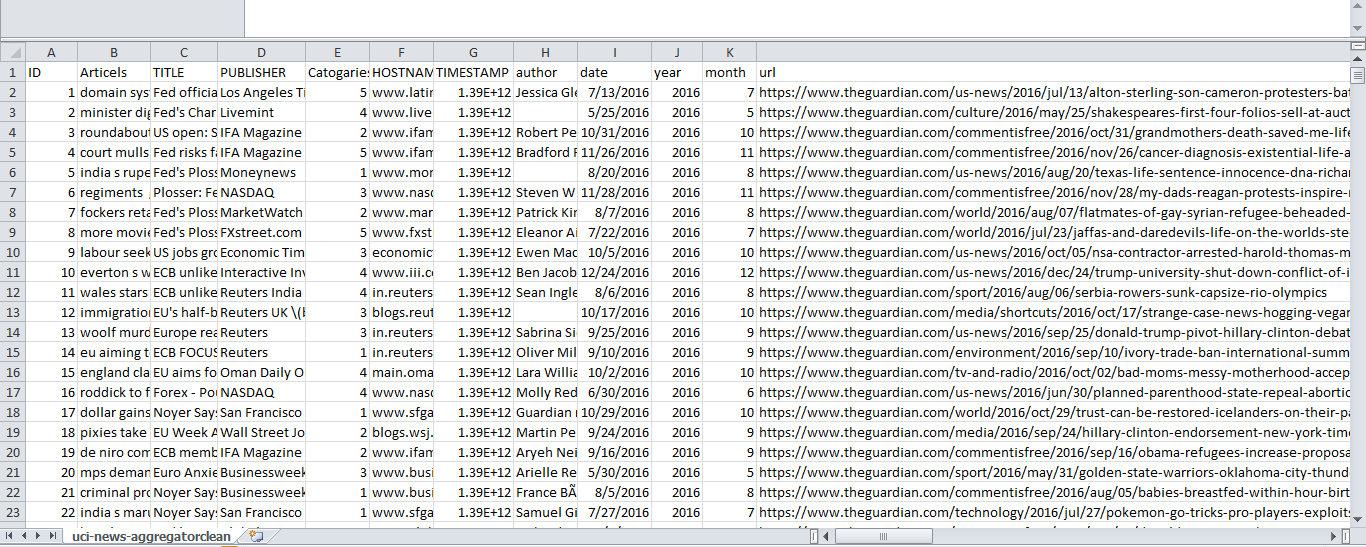
14

**4.1. Data Acquisition:**

A first step in Text Mining Process and it’s responsible for:

* **Collecting Data from many different Resources:** we collect the Data from kaggle and Razan Sent a message to a Reuters News agency and she contact them until they gave her a large number of data approximately 6000 records.
* **Cleaning Data:**

We use a RapidMiner tool to clean the Data and change it in a way that will be easy to understand and deal with it.

This picture shows how the data was after collected and before we cleaning it:

**At the begging we combine the title and Articles columns and put them in a one new Column in Excel.**

15

**Then we load Data on the RapidMiner Using a Read CSV process and then we use a cleaning process to:**

* Remove Duplicate
* Remove a numbers and URL and special character’s such as $#%^&\*
* Remove Missing value
* Remove Extra spaces

**Now we will write in more details about these steps.**

**Numerical to polynomial process:** The Numerical to Polynomial operator is used for changing the type of Numeric attributes to a polynomial type.

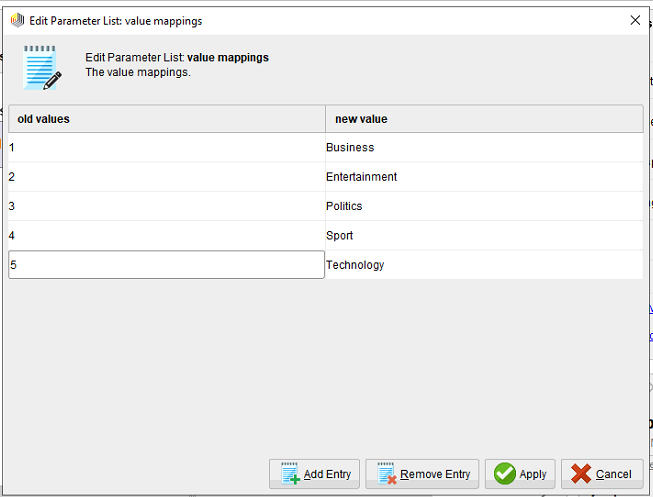


Because the Categories was written in a numbers from 1 to 5 so we change The type to polynomial to start put a name of Categories instead of number using \* **Map Process** .

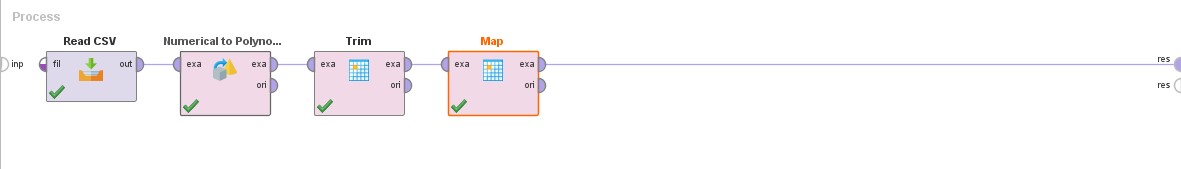
**\*Map process:** Specified values of selected attributes to new values. This operator can be applied on both numerical and nominal attributes.

16

18

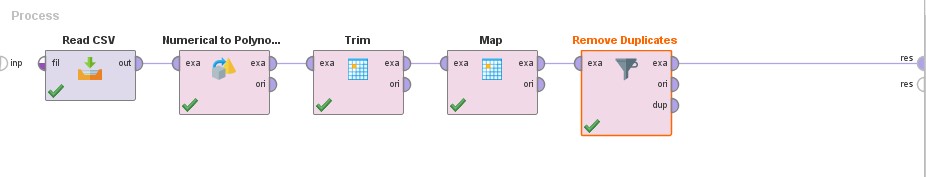


**Trim process:** This process help to delete and remove a spaces that we don’t needed.



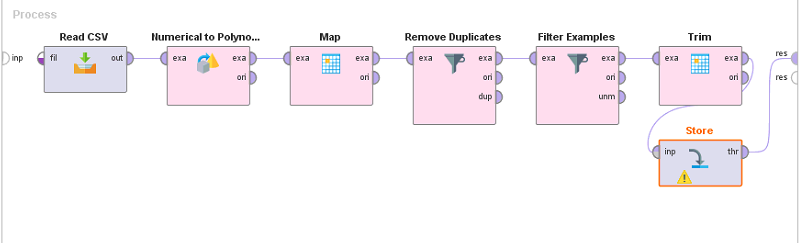
Then we delete the duplicated rows using **Remove duplicate Process** then we only want the Articles and Categories column so we use **Select Attribute Process to choose this 2 column.**

17

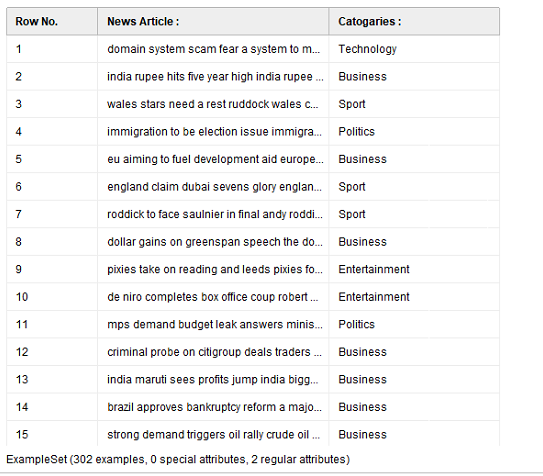


Then we use **a Filter Example Process** to remove a missing value, numbers extra Characters’ and URL.

Then we store the process in Repository as a data and a process to use it again using **Store Process.**

****

18

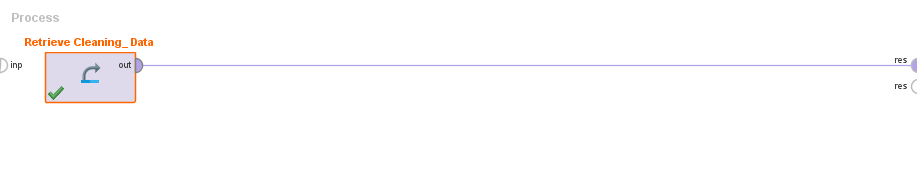
**The final shape of Data after Cleaning:**

**Here is the visualization of the our Data:**

19

**4.2-Text preprocessing:**

Is an important task and critical step in Text mining, In the area of Text Mining, data preprocessing used for extracting interesting and non-trivial and knowledge from unstructured text data.

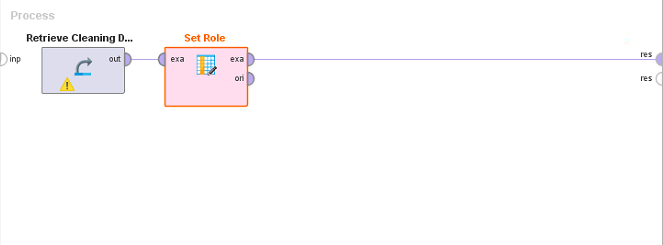
1. We retrieve the data which is cleaned now using a **Retrieve Process** This Operator can access stored information in the Repository and load them into the Process
2. We put a categories as a **label** because we want to use these classified class to build a Machine Learning then the result of new document is to predict in the category of document using **Set Role Process.**

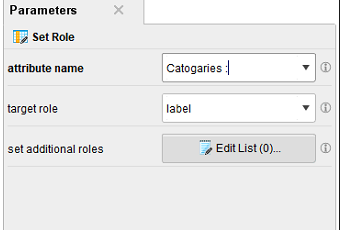
**Set role Process:** an Operator used to change the role of one or more Attributes.

**Parameters of this Process:**

* **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.

20

* **Label:** This is a special role. An attribute with the label role acts as a target Attribute for learning Operators. This label is also often called ‘target variable’ or ‘classes.

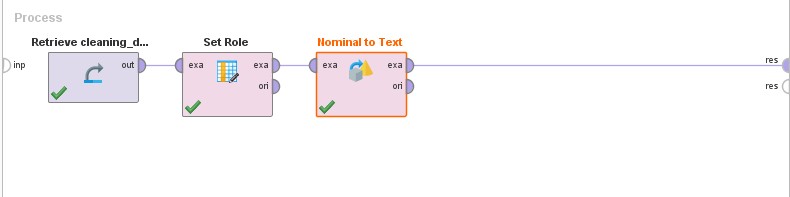


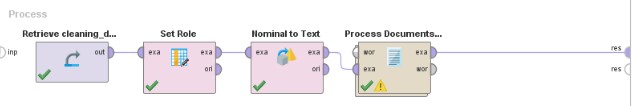
23

1. Then we use a **Nominal to text Process :** Is to specify which column is a text column, since RapidMiner “Process Document ” Operators work only on text data.

21

1. Preprocessing step using **Process Document from Data Operator :**

Which is an operator used to create word vectors from text attributes, this is a nested operator that contain sub operators inside it.



And we use **TF-IDF** as a base to train our model, which 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.

**Preprocessing steps:**

* **The tokenize operator** tokenizes documents, and we select in the parameters of this operators 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 operators splits the text of a document into a sequence of tokens).

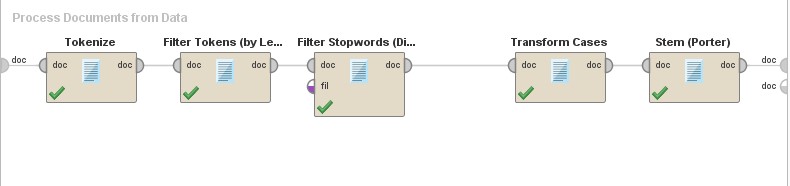
22

* **The Filter Stop words (Dictionary)** operator applies stop word list from a file. Stop words are words which are filtering out prior to, or after, processing for natural language data (text) For example, some of the most common stop words for search machines are: the, is, at, which, and on.

**We collect a stop word from a WorldNet Dictionary.**

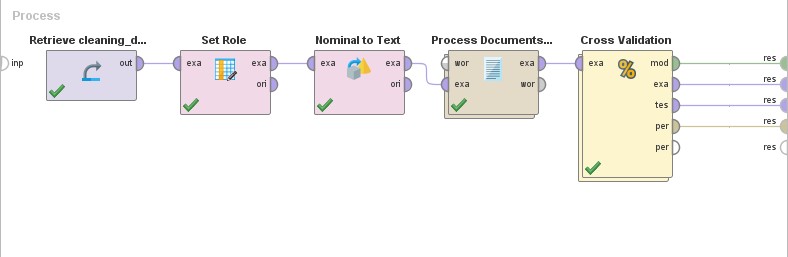
* **The Filter Tokens (by Length) operator:** filters tokens based on their length. In its parameters we select the min chars of a token to be 3(thus removing single letter words), and the max chars of a token to be 20 which is safe enough to say that words consisting of 20 characters are probably gibberish.
* **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 from or steam e.g. like, liking, likely, unlike belong to like.
* **The Transform cases operators** transforms the cases off all characters. In its parameters we choose to transform all characters to lower Case.

23



**4.3- Text Modeling :**

The process of selecting the model that we will use to learn the algorithm from training data. This will be after Text transformation, features extraction and features selection.

**Here we use a Cross Validation Process with**:

We deal with a Number of train data sets =300 row (text, categories)

K (number of folds) =10

24

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Train | Train | Train | Train | Train | Train | Train | Train | Train |

**Best learning result**

**Best validation**

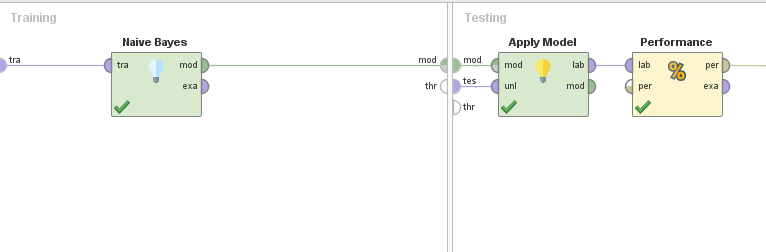
The Cross Validation operators is a nested operator. It has two sub processes. A Training sub process and a Testing sub process. The Training sub process is used for training a model. The trained model is then applied in the Testing sub process. The performance of the model is measured during the testing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

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 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.

25

And this process contains sub process as shown below:



2

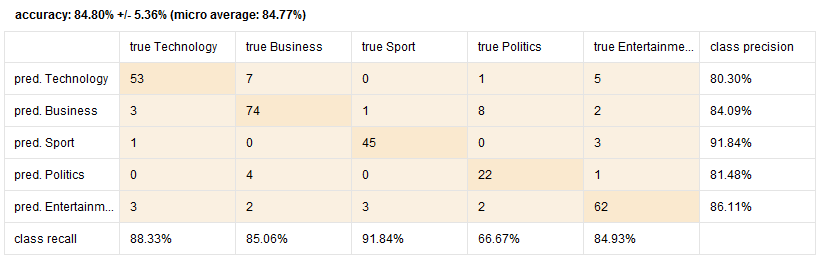
1

**Number 1 is the (training part) that** contains the Classification model (Nearest neighbor).

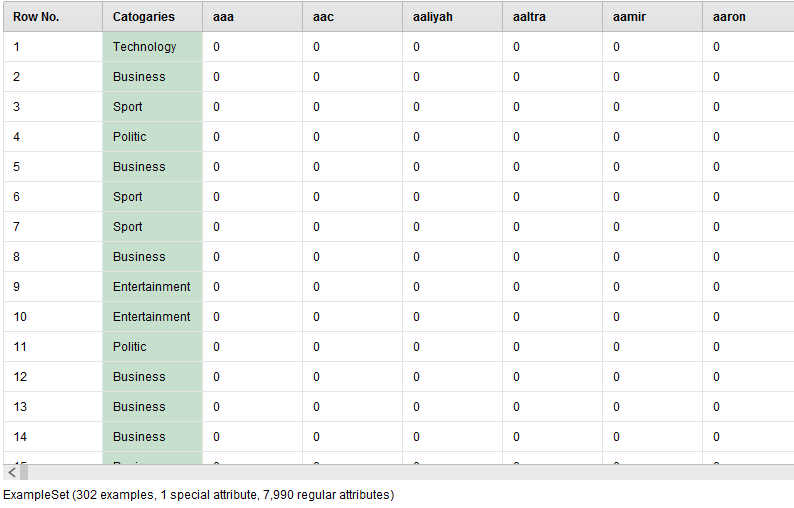
**Number 2 is the (testing part) that contains:-**

1. **The apply model** (The 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).

26

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

After that we build the model (with different operators) that described in detail in the previous pages (in training data for 5 classes), and run the process to show the results below:

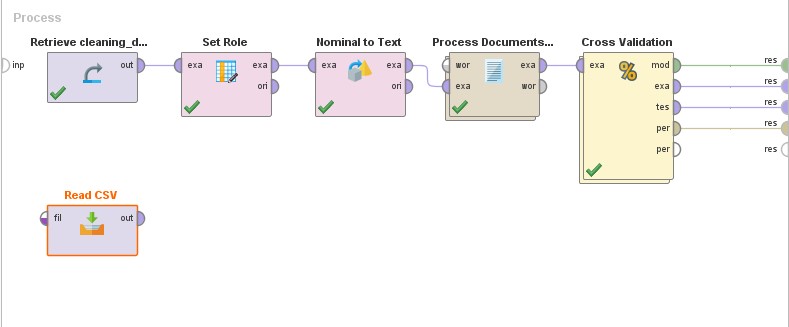


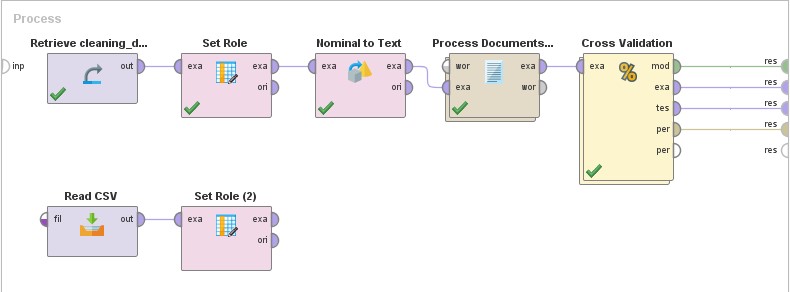
27

**4.4- Testing and Evaluation**:

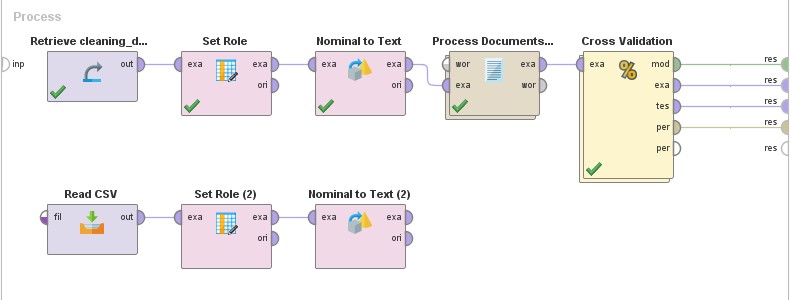
The Creation of testing data using a following process:

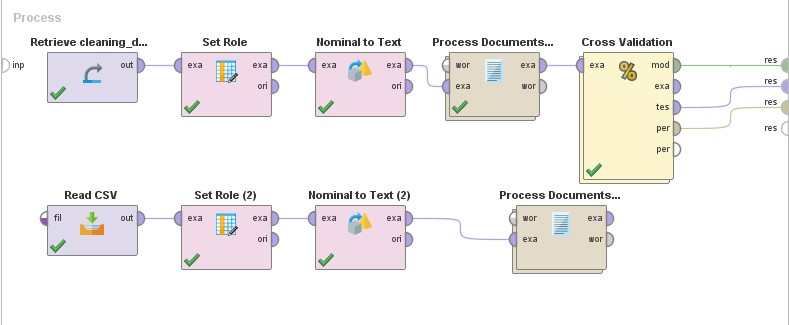
1. We use the Read CSV process to load News that need classifications a test Data.



1. **Set Role** process again as we made in the training data.

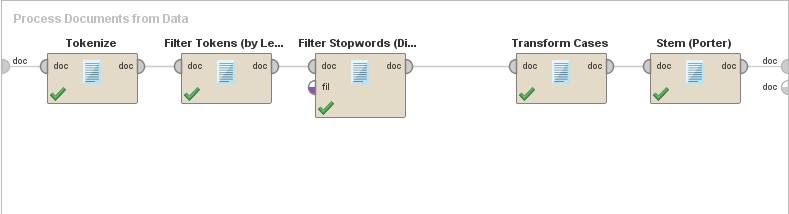
28

1. We use the **Nominal to text process** again as we made in the training data as we explained before.
2. We use the (**Process documents from data**) that we explain it previously to clean the text using a lot of sub processes inside of it as we will see below:

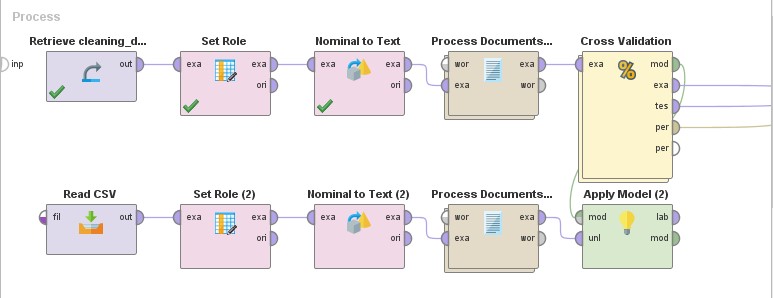


29

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

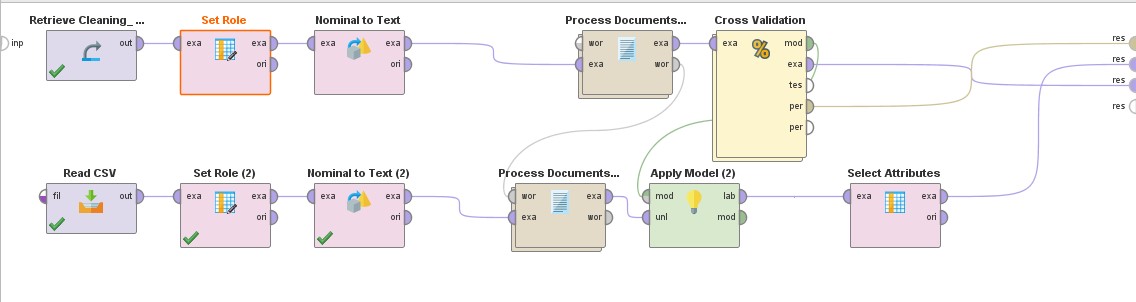


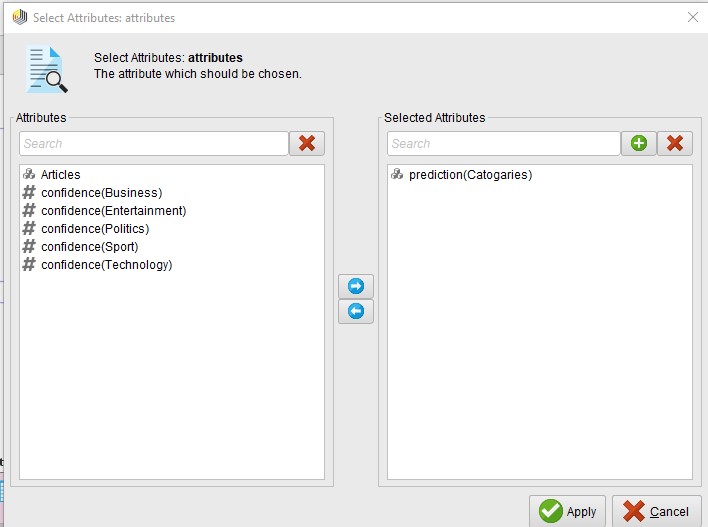
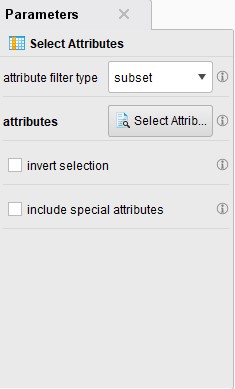
5-We execute the model using the **Apply model process,** this operator applies a model on an Example Set. As shown below:



1. **Select Attributes Process** : The Operator provides different filter types to make Attribute selection easy

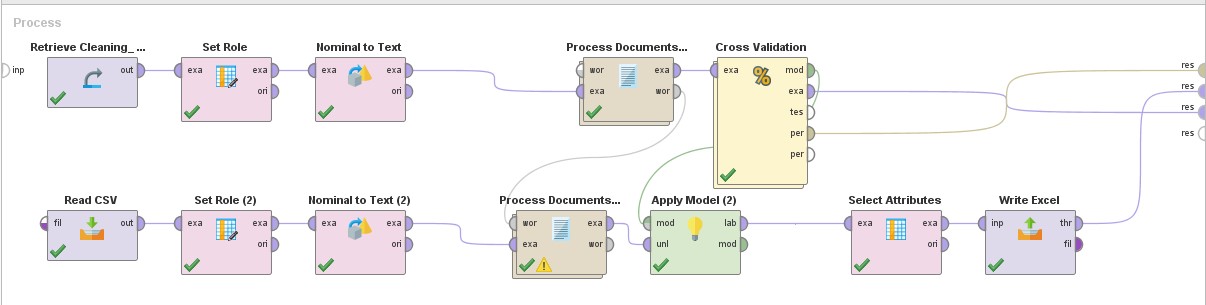
30

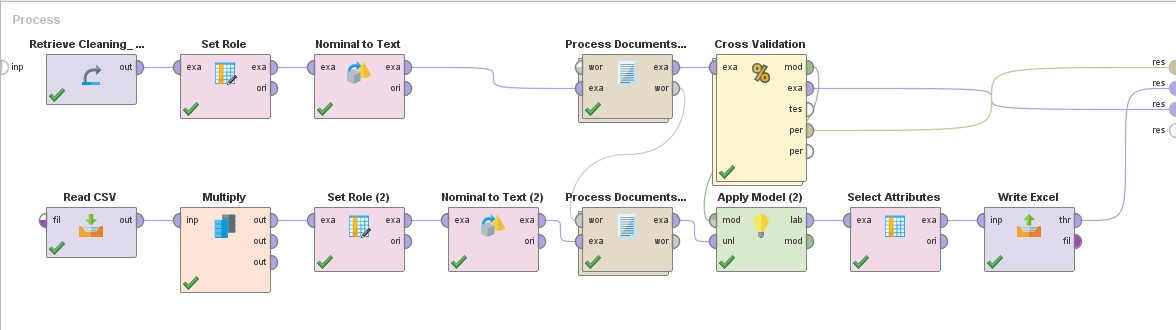




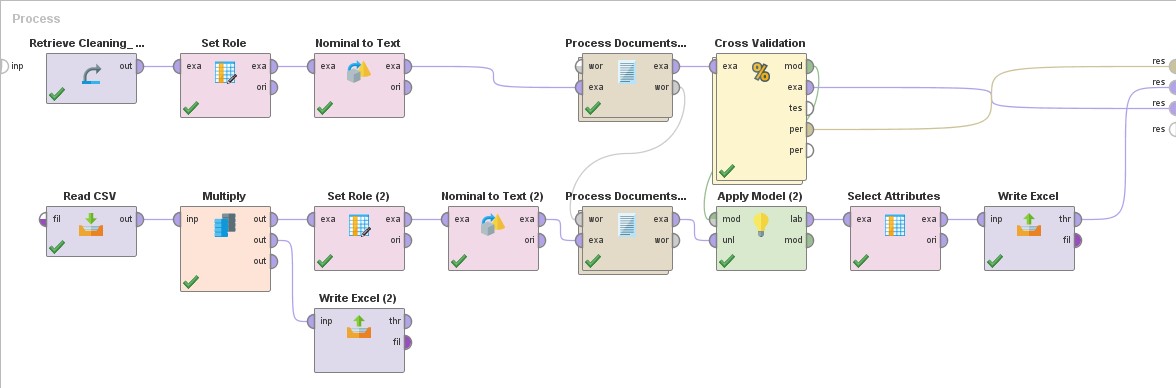
1. And then we must take all the predicted news that done by the RapidMiner using the **Write excel** process as follows:

31

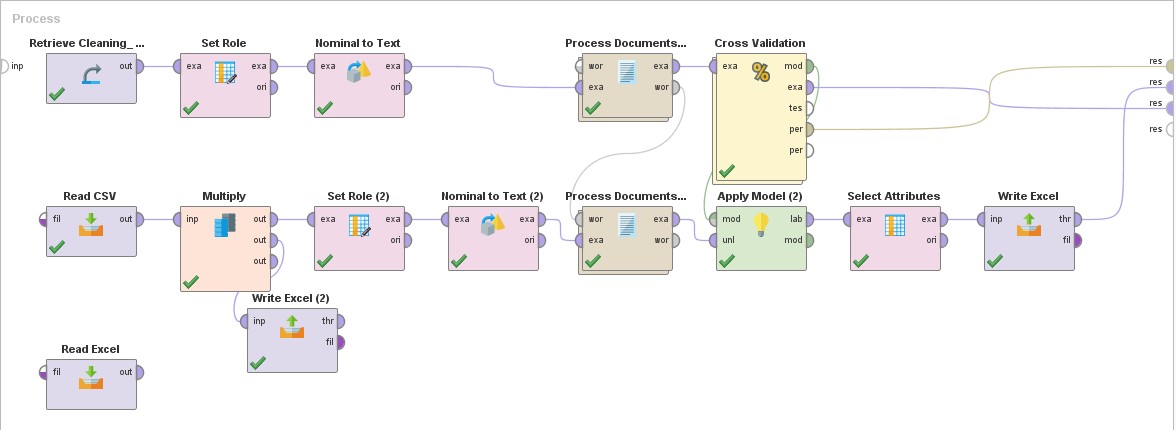


1. Many Operators have an output port named *original* or *throughput*, which does not change the input. By chaining Operators to one or more original ports also copies of an object can be created. Use **Multiply Process**
2. we use the **Write Excel** process again to store 2 Column (Articles and Categories )

32

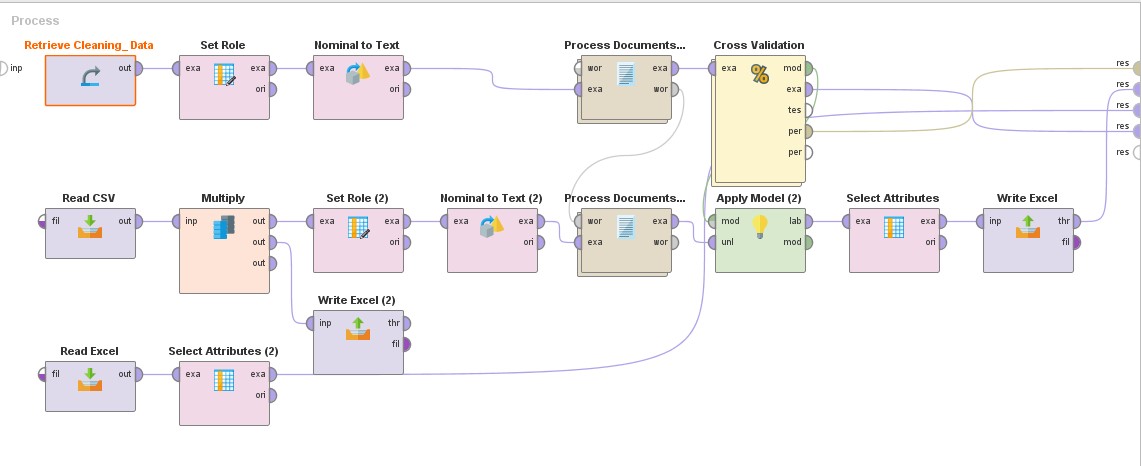


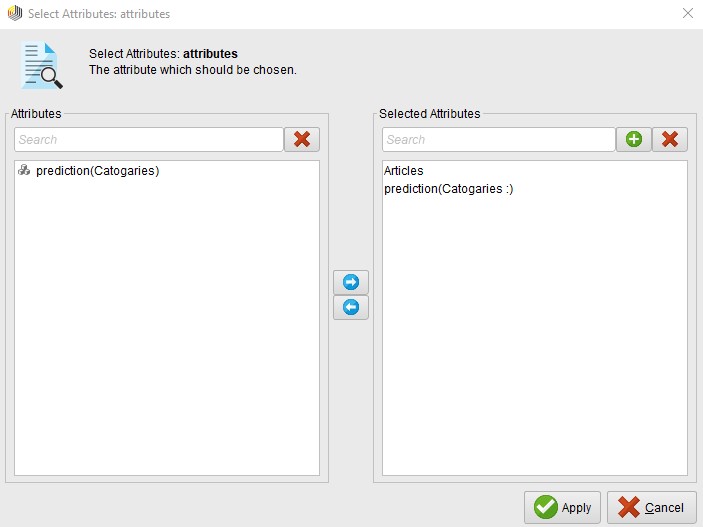
1. **Read Excel:** is contain data predication.



1. We use the **Select Attributes** process again.

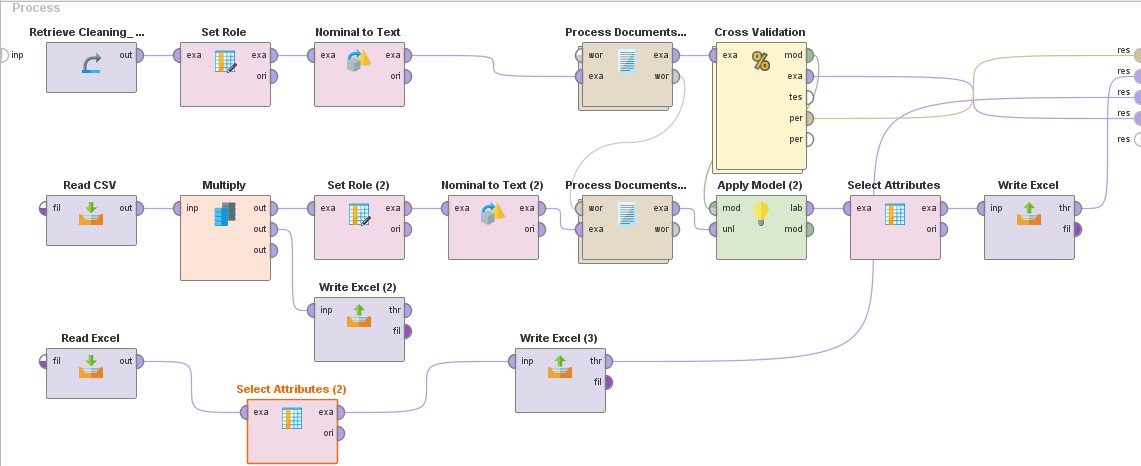
33





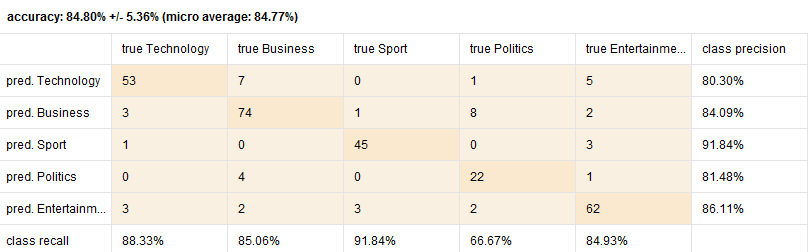
34

We use the **Write Excel** (3) process again to obtain only the Articles and Categories which display to the organizations as a final result.



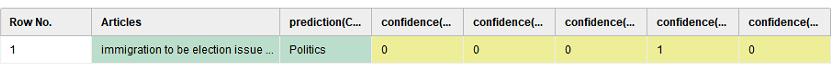
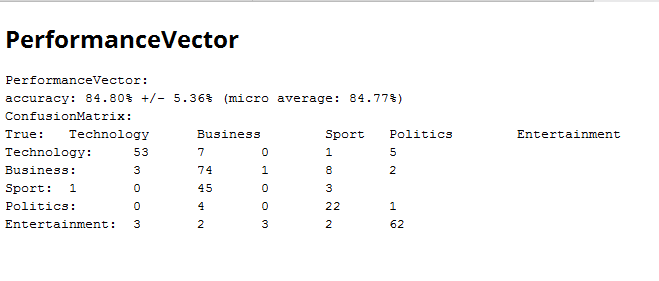
**The Final Result :**

We see is as follows after we **click run:**



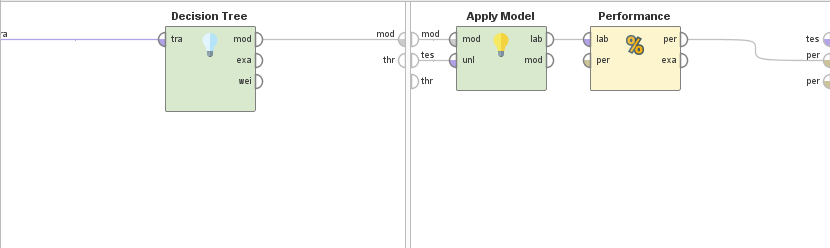
We show the ratio of correct predictions to total predictions 84.80% that is classification Accuracy.

35



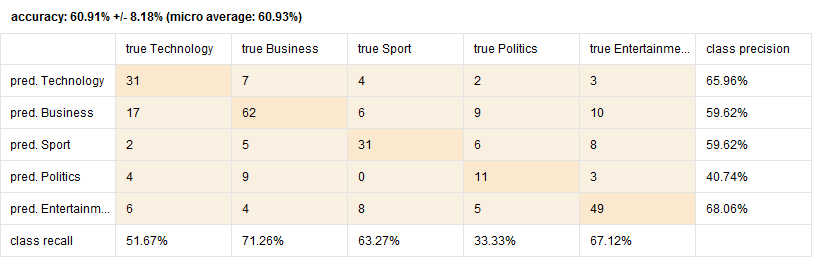
Now we want to try a different classification model to see if the Accuracy and Time change (increase, decrease) to decide which classifier is better to use.

**First**: we want to choose **Decision Tree** classifier

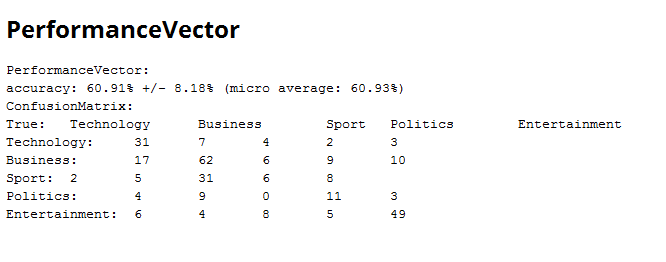


The **confusion matrix** is as shown below:

36

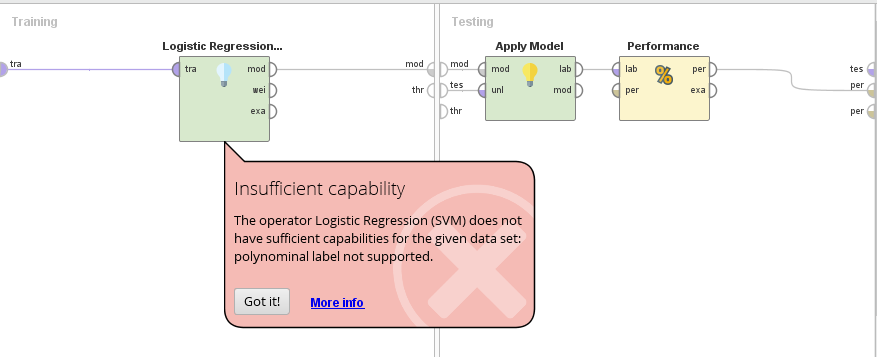


**The Confusion Matrix:** is a technique for summarizing the performance of a classification algorithm . And calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.



**Second:** we use **SVM** (Support Vector Machine) algorithm :

36



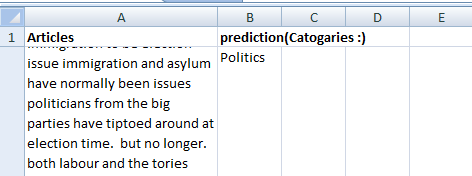
As we see the error, SVM does not support polynomial (Entertainment, Politics, Business, Sport, and Technology), it’s only support binomial.

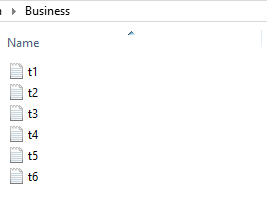
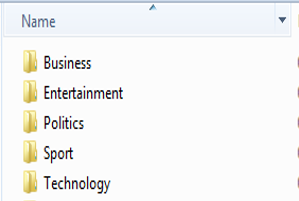
|  |  |  |  |
| --- | --- | --- | --- |
| Model | Naïve bays | Decision Tree | SVM |
| Accuracy | 84.80% | 60.91% | Does not support |
| Time | 3s | 6s | ----------------------- |

**So according to this result we recommend the company when they want to classify use a naïve Base as an algorithm for building Machine Learning .**

37

* Then when the Result is reach an Excel File it’s look like that :

****

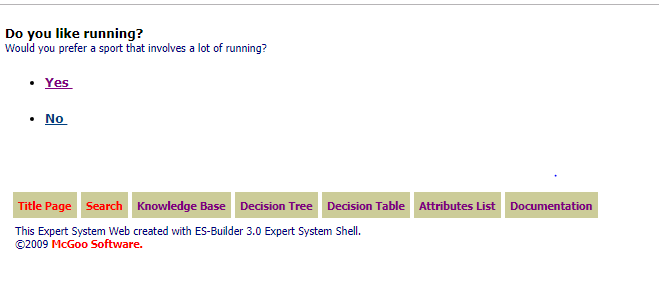
* ****Finally we Transform the excel Data to many text files using a Kutools for Excel tool and each of these text has a defined Categories.

38

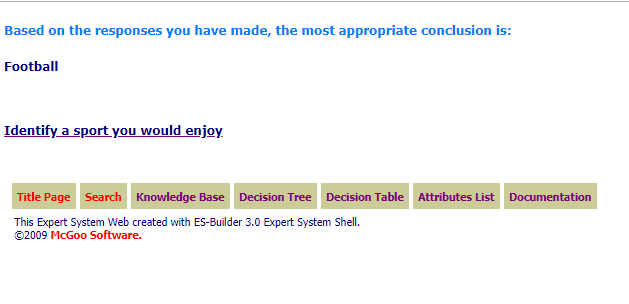
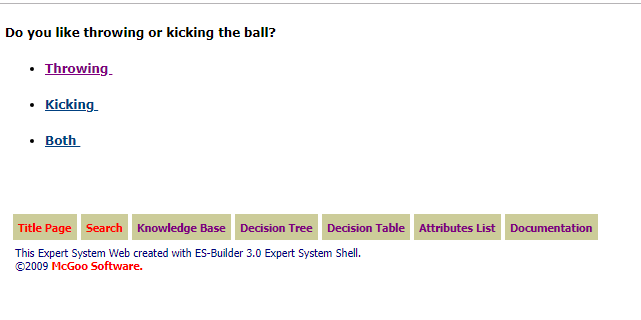
* 1. **Application :**

**Text Mining applies in a variety of areas such as Web Mining, Medical and Resume Filtering.**

In our project we use a Corvid Expert System in order to build system that can help the organization to know what consumers prefer and like then when the consumer go back he can easily find what he want by asking several question here we will focused only on one part which is sport :

****

39

****After answering these question this the result then the organization conclude that he prefer these types of news and display to costumer once entering the site**.**

40

1. **Problems and Recommendation :**

**5.1 Problems that faces us during our project while Dealing with RapidMiner :**

* Tool not fully support some languages such as Arabic.
* Tool does not support the huge number of data.
* Tool must be paid to give better results in pulling data.
* Lack of sufficient resources to learn through.
* Tool takes a lot of time to process the documents and because our computers as a students does not have a high processer so we face this problem when we dealing with a huge number of Data.
* Tool can’t predict to more than one column (Hierarchical classification ( .

**5.2 Recommendations:**

Acording to the Result that we obtained we recommend the companies to use our system with a RapidMiner Tool to classify their Documents instead of Manual Classification and use a Naïve Base algorithm for Doing that , because this will help to reduce the Employees time and efforts and make their work done more effectively and efficiency .

41