Hands-on Tutorial on Sentiment Analysis using KNIME

Raghava Mukkamala (rrm.digi@cbs.dk)

This hands-on exercise will use the KNIME Analytics Platform to build a sentiment analysis KNIME model on the textual data from online discussion forums about diabetes. As part of the workflow, we will use the *Palladian Text Classifier,* which uses a supervised machine-learning approach to the discussion texts, to predict the sentiment of the discussion, e.g., positive, negative, or neutral.

Prerequisites

Before starting this exercise, you should have installed the KNIME Analytics Platform on your computer.

If you have not installed the KNIME Analytics Platform on your computer, then please go through the following URL to download. https://www.knime.com/downloads.

Here is the installation guide to KNIME:
https://docs.knime.com/latest/analytics_platform_installation_guide/index.html#_installing_knime_analytics_platform

1 Dataset: Online Discussion Forums about Type-2 Diabetes

Before we begin creating a KNIME workflow to build a sentiment model, we must understand the diabetes discussion forums dataset first. The online diabetes discussion forums dataset is provided to you as part of this hands-on session in the form of an excel file. Therefore, we will first explore the attributes/column names of the dataset displayed in the following table.

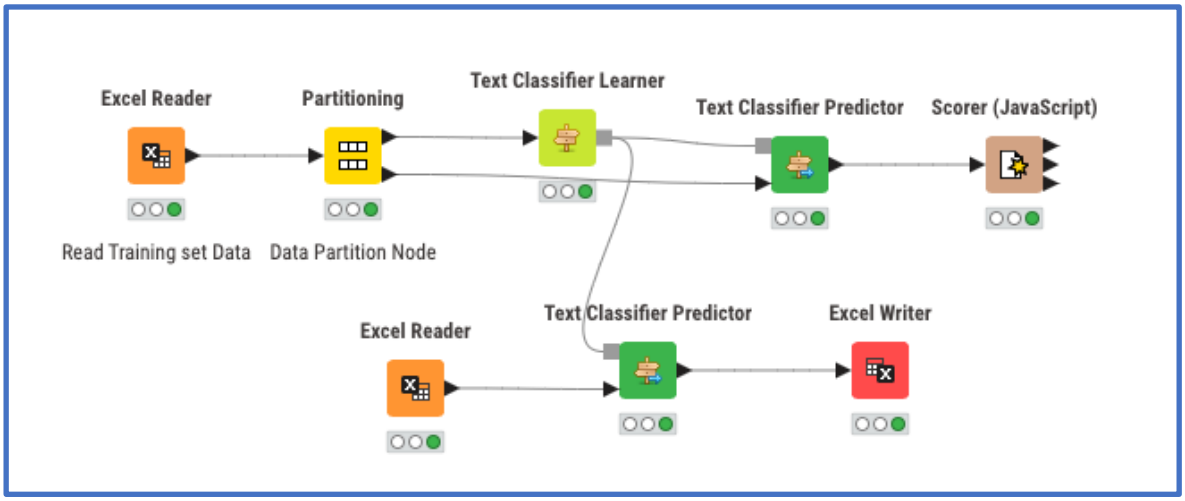
Data Attribute	Explanation
No	Serial number of discussion text
Discussion_text	Text from the discussion forum for 1 thread
Label	Sentiment Label: positive, negative, neutral. (**dependent variable**)
ModelName	Name of the model. In this case, since we are building sentiment analysis, the value is the 'Sentiment'.

The first few records in the dataset are shown below.

	No	Discussion_text	Label	ModelName	TextId
2	1	...been on a low-carb diet my blood glucose has been pretty normal although I did have hypo-type symptoms when I briefly tried taking ...associated .... Type 2 diabetes it's something I am always...	Neutral	Sentiment	508e98fc-8db4-4283-a2a4-63c3b70acf86
3	2	A major problem with the American diet is too much refined grains and added sugar which are associated with the rise in obesity and type 2 diabetes [Source](http://www.joslin.org/news/Research-affirms-good-nutrition-can-help-prevent-and-control-type2-diabetes.html). This article is from the results of a systematic review of randomized controlled studies which represents high-tier evidence.. Edit: I'll add wheat bread can be great and is usually very high in fiber (which is absolutely critical to digestive health).. Potatoes have very few redeeming qualities (high in potassium and a few minerals are the only ones I can think of).	Negative	Sentiment	1349404b-4192-4da2-89ed-346d17351c6b
4	3	A major problem with the American diet is too much refined grains and added sugar which are associated with the rise in obesity and type 2 diabetes [Source](http://www.joslin.org/news/Research-affirms-good-nutrition-can-help-prevent-and-control-type2-diabetes.html). This article is from the results of a systematic review of randomized controlled studies which represents high-tier evidence.. Edit: I'll add wheat bread can be great and is usually very high in fiber (which is absolutely critical to digestive health).. Potatoes have very few redeeming qualities (high in potassium and a few minerals are the only ones I can think of).	Negative	Sentiment	1349404b-4192-4da2-89ed-346d17351c6b

2 Build KNIME workflow for Sentiment Analysis

After completing the hands-on exercise, the final KNIME workflow will appear, as shown in the following figure.



You will build a KNIME workflow to build a sentiment analysis model using the Palladian text classifier on the diabetes discussion forums dataset to identify the sentiment of the online discussion. The process involves the following steps.

1. Reading and preparing the data
2. Build Sentiment Model
3. Make predictions for the sentiment of new discussions

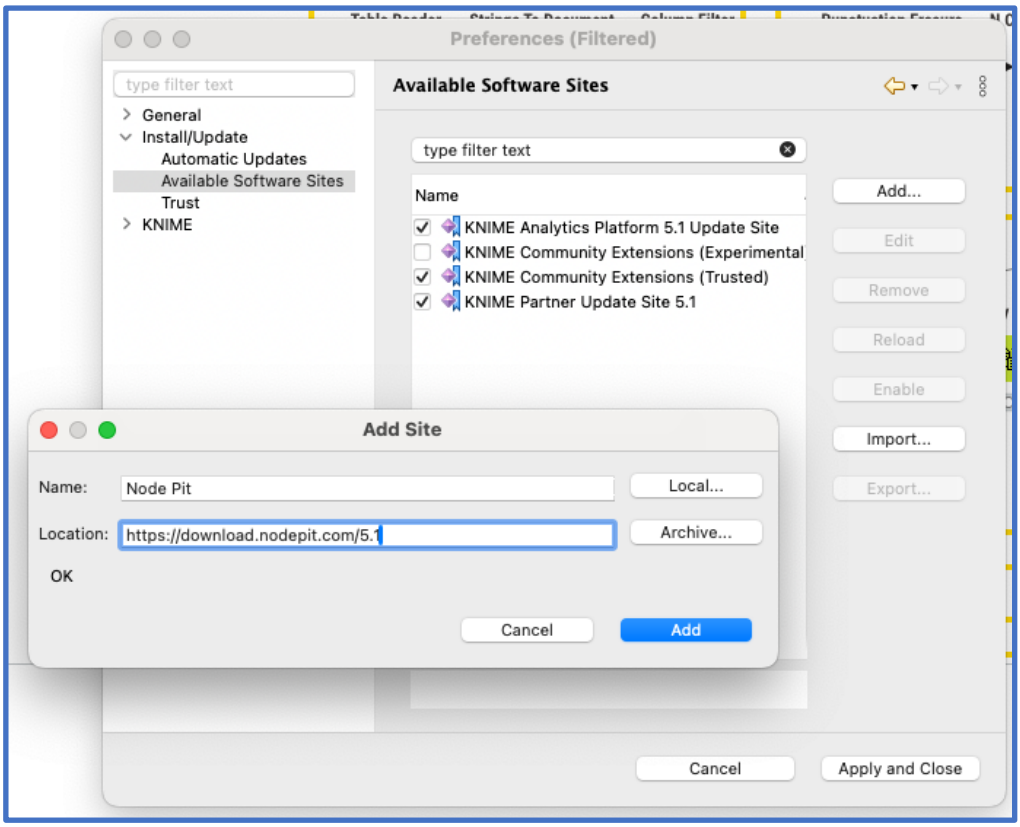
We will go through each of these steps in more detail below. Before you start, if you have not

watched previously, we recommend that you watch the video (only 2 minutes duration) named “What is a Node? What is a Workflow?” (https://www.youtube.com/watch?v=4rETNe-Xx7k) one more time to recap the basics of how to build and execute KNIME Workflows. Before developing the KNIME workflow model for sentiment analysis, we need to install the Palladian Text Classifier node from the Nodepit KNIME repository.

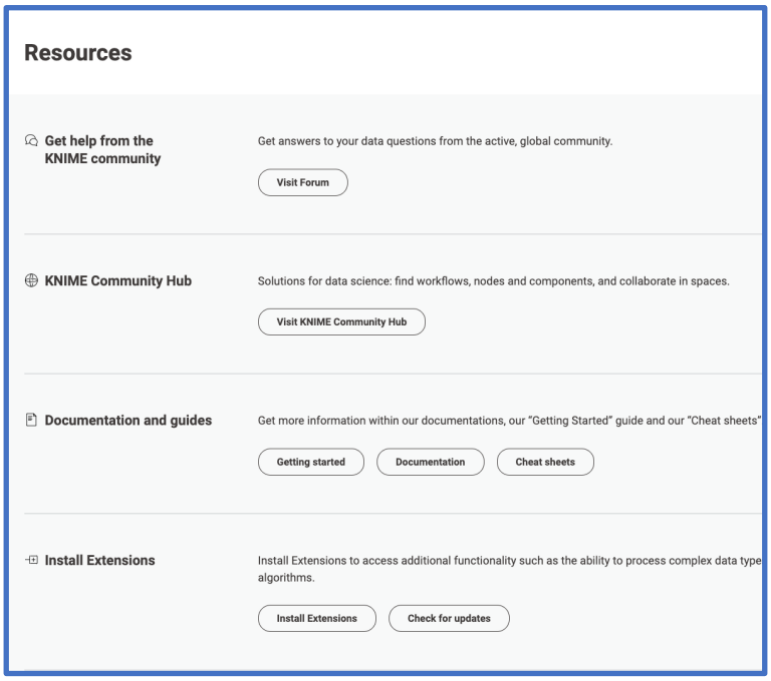
2.1 Installing Palladian Text Classifier

To install the Palladian Text Classifier Node from the Nodepit repository, first, you need to install NodePit for KNIME analytics platform by following the instructions specified at: https://nodepit.com/product/nodepit/installation.

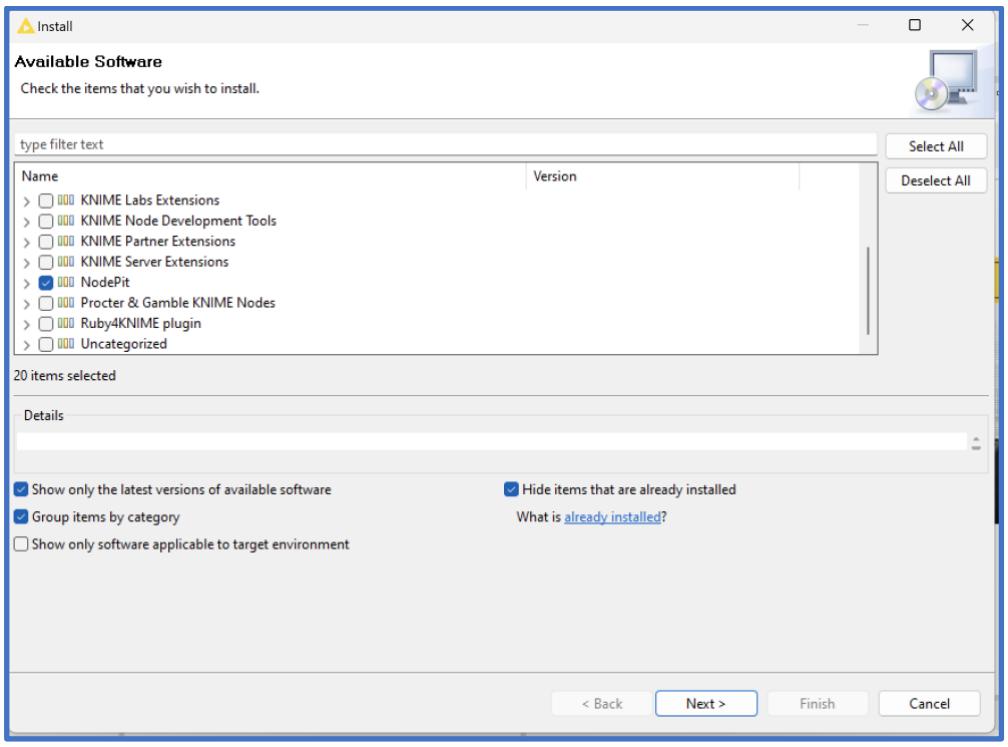
First, you need to add the URL (https://download.nodepit.com/5.2) to the NodePit repository in the Available software sites. In the KNIME platform, click the settings icon  on the right-hand side of the KNIME platform, go to *Install/Update* → *Available Software Sites*, click on *Add*… and paste the update site in the location as shown in the following screenshot. Make sure the update site matches your current KNIME version, e.g. https://download.nodepit.com/5.1. Please note that the URL to the Nodepit repository is https://download.nodepit.com/5.1 (no ending period).

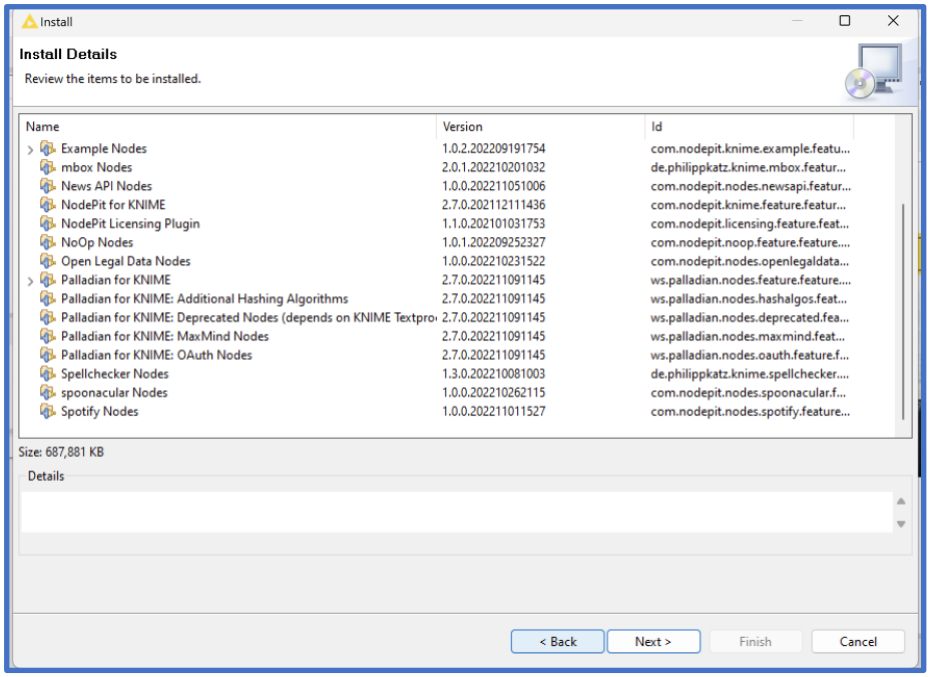


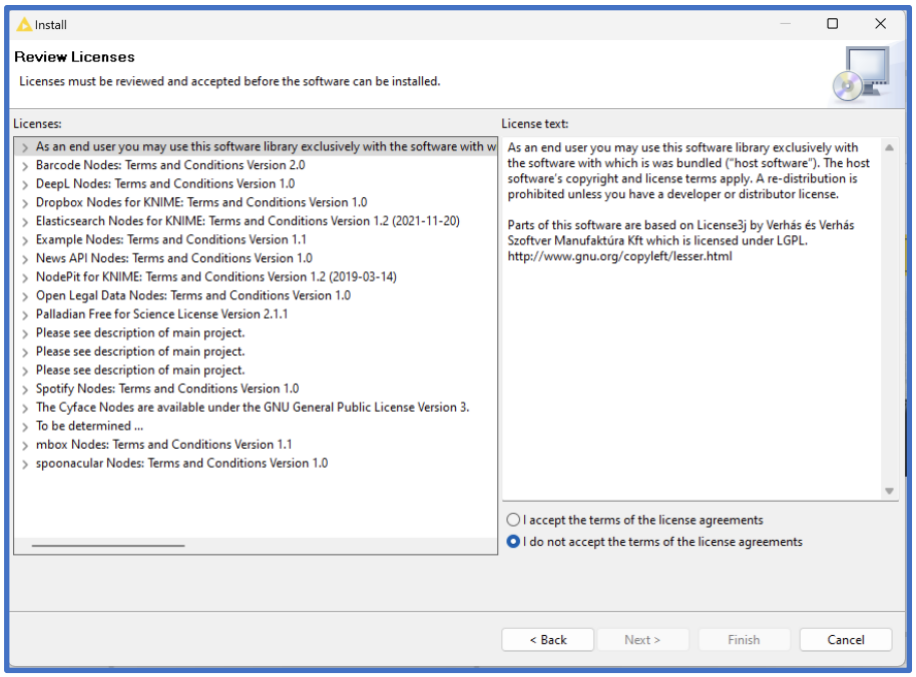
After adding Nodepit to available software sites, again click the **i** icon () and click on the ***Install Extensions*** sub-menu in Resources, as shown in the following screenshot.



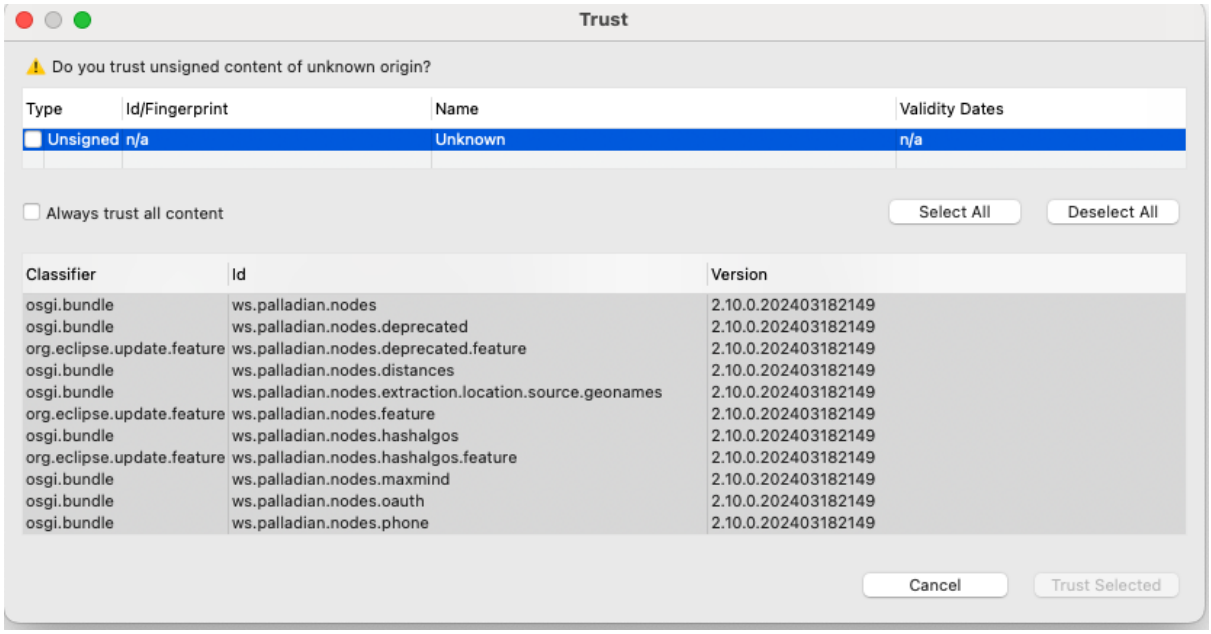
It will open the install pop-up, as shown below. Select the *NodePit* in the list of available software extensions and click *Next* to proceed with the installation as shown in the next five screenshots.

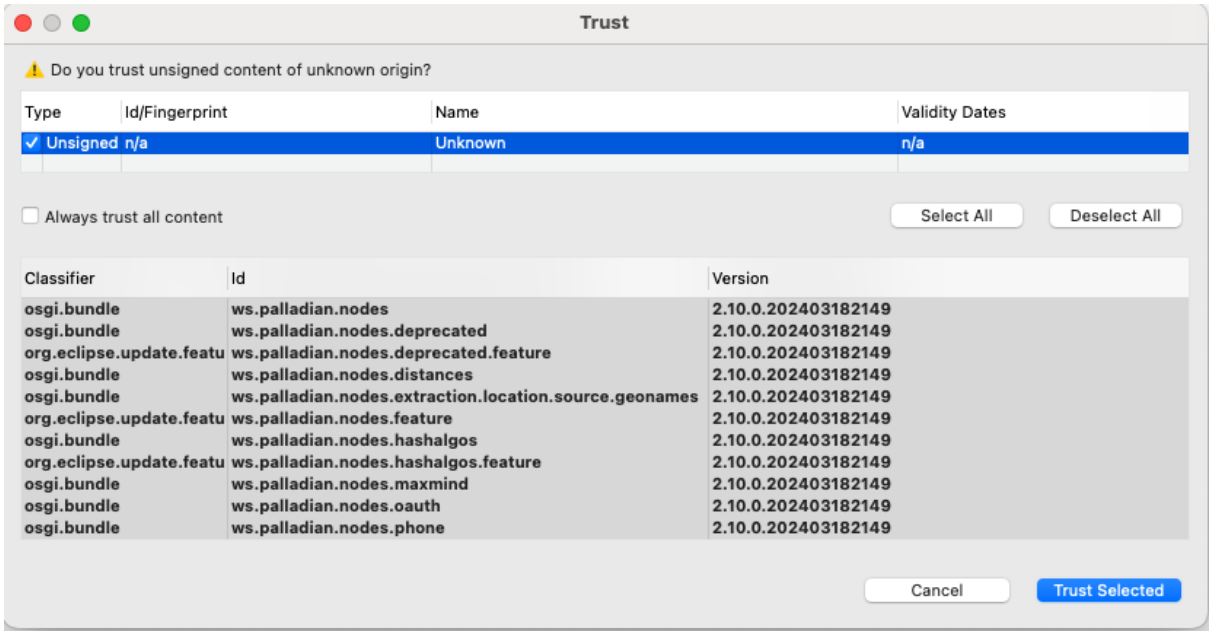




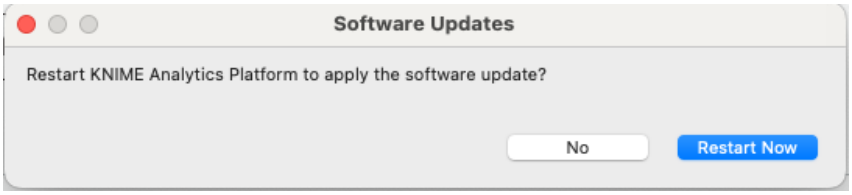


It will complain that the Nodepit software is not signed or is from an unknown origin. To install the nodes from Node Pit, you need to check unsigned n/a and proceed to install them by clicking *Trust Selected*.



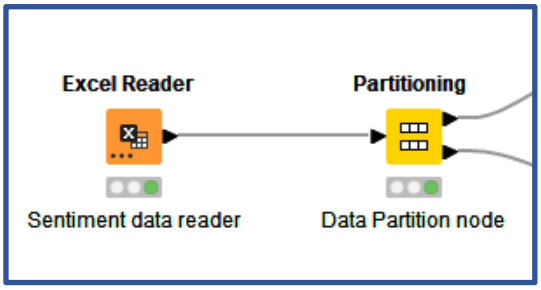


After finishing installing the NodePit extensions, restart the KNIME analytics platform to load the new extensions.



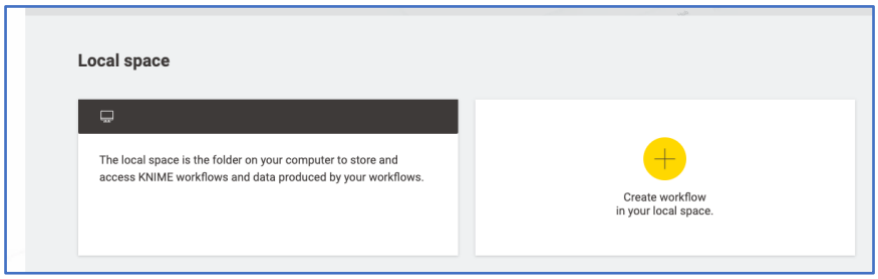
2.2 Getting and preparing the data

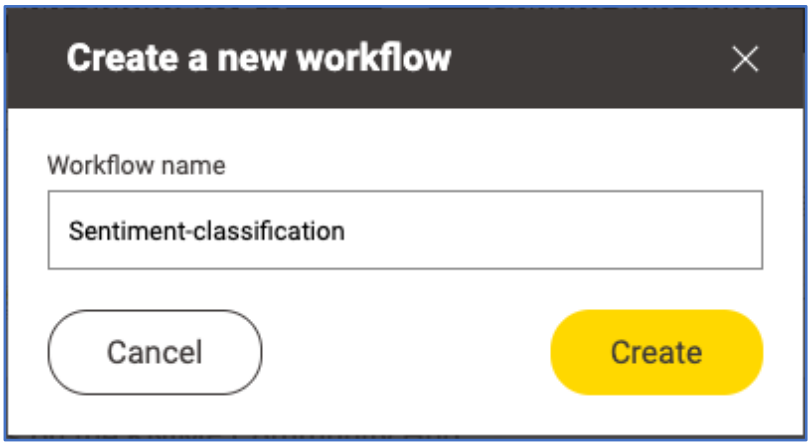
This step involves two nodes, as shown in the following picture.

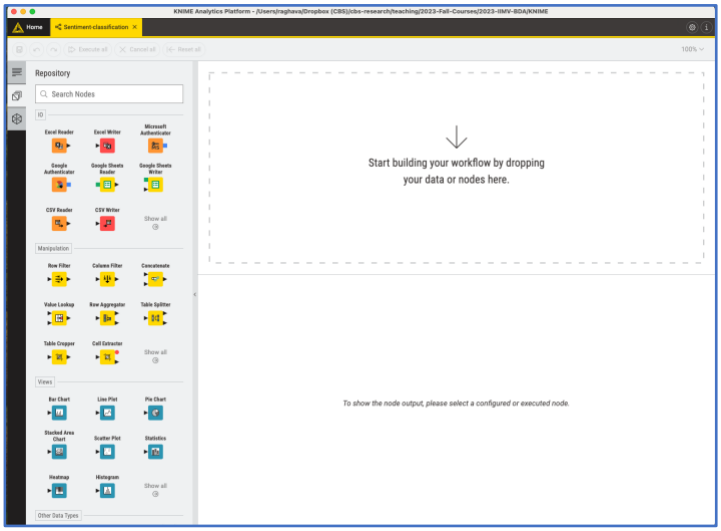


First, you should download the online diabetes discussion forums dataset (*discussion-forums-sentiment-training-set-data.xlsx*) from the canvas, which is published under the Datasets heading for this hands-on session. Download the dataset to your computer (if you have not done that already) and make sure that you know the exact location of the file on your computer so that you can later open the file in KNIME. On Windows, you use the *File Explorer* program to browse folders and files. On Mac, you can use *Finder*.

When you have downloaded the dataset file, then open the KNIME Analytics Platform, right-click your local workspace in the KNIME Explorer, and select New KNIME Workflow…. Select a name for the workflow and click Finish as shown in the following screenshots.



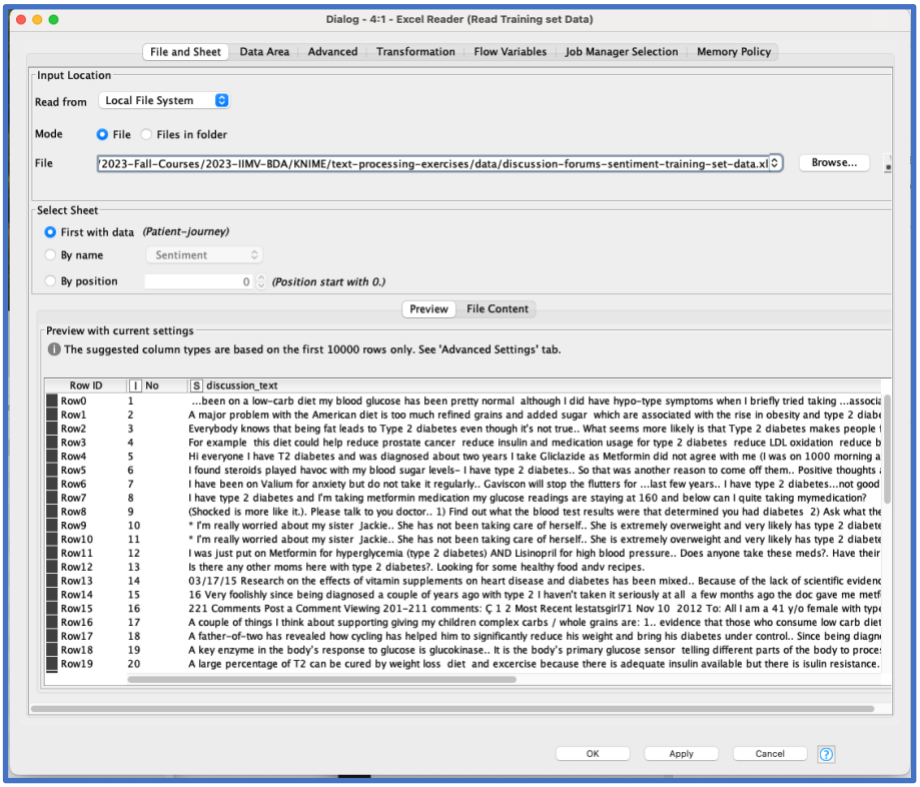




2.2.1 Excel Reader Node

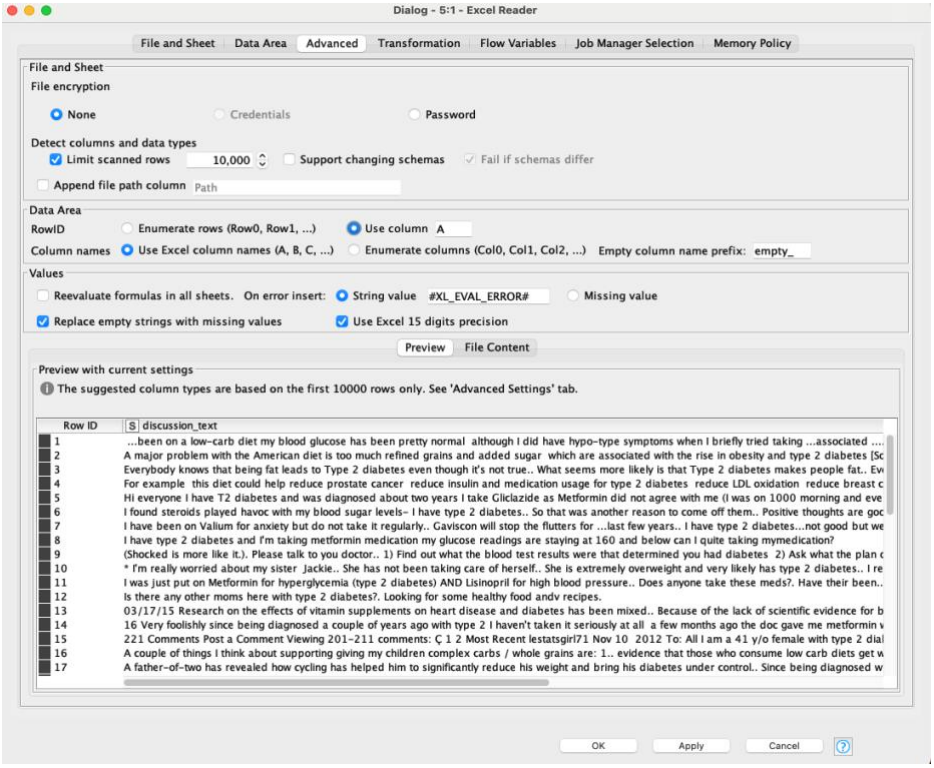
Since the dataset is an excel file, we will use an ***Excel Reader*** node to read the data. You can find an ***Excel Reader*** in the Repository by typing "*Excel* Reader" in the search field.

Drag an ***Excel Reader*** node to the Workflow Editor. Then double-click the ***Excel Reader*** to open its configuration dialog. In the configuration dialog, click the "browse…" button and try to find the .xlsx file (*discussion-forums-sentiment-training-set-data.xlsx*) containing the dataset. Please make sure to select the sheet name as **Sentiment**, as shown below.



Then click the Advanced tab and indicate that the file contains both RowId and

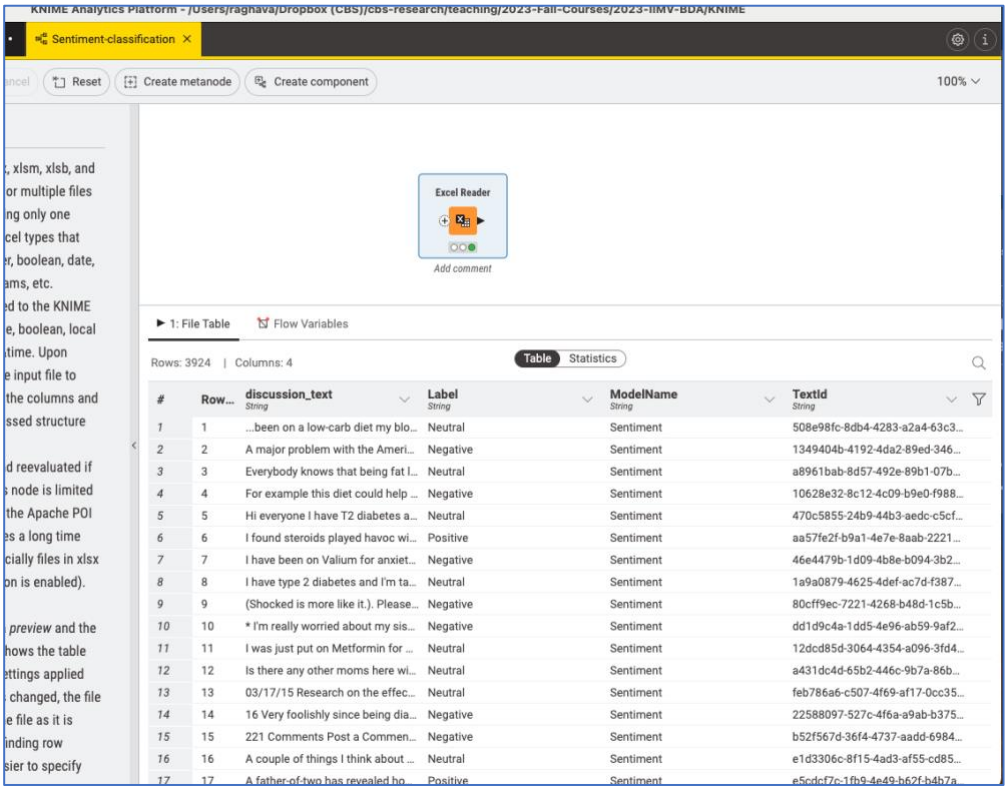
Column names indicate that as shown below.



If you cannot find the dataset file from the File Reader configuration dialog, use the following approach instead:

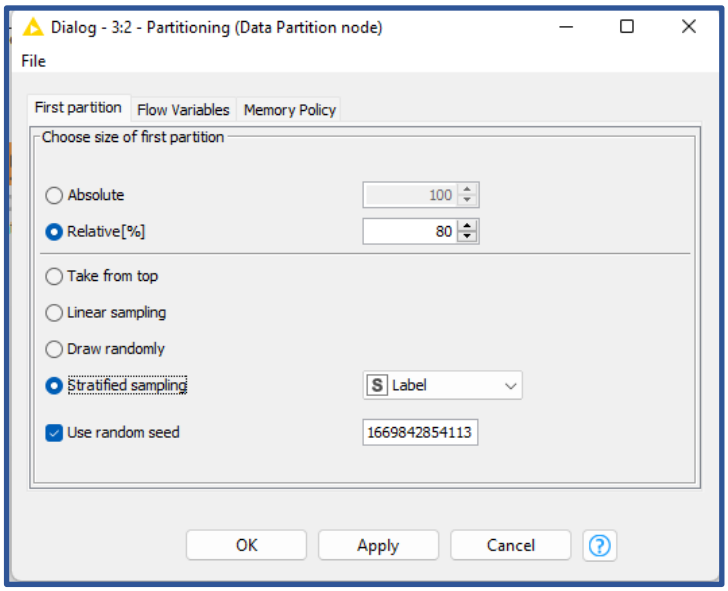
1. Open File Explorer if you are using Windows or Finder if you are using Mac.
2. Open the Downloads folder or the folder where you have downloaded the files.
3. This folder should contain the dataset file you downloaded from the course page.
4. Drag the dataset file and drop it on the KNIME Workflow Editor. This will automatically create a File Reader that is configured to read the dataset file that you dragged and dropped on the Workflow Editor.
5. Make sure that the settings are assigned as shown in the above figure.

When you have configured the excel Reader, then right-click it and select Execute. When you click the node, it will show data from the Excel file in the File Table below as follows.

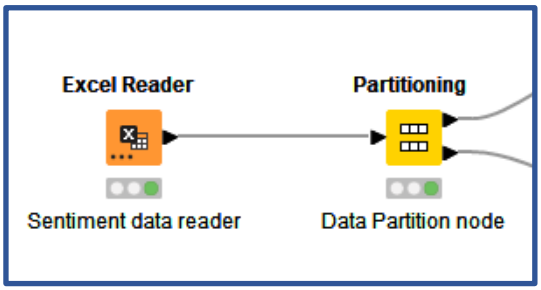


2.2.2 Partitioning Node:

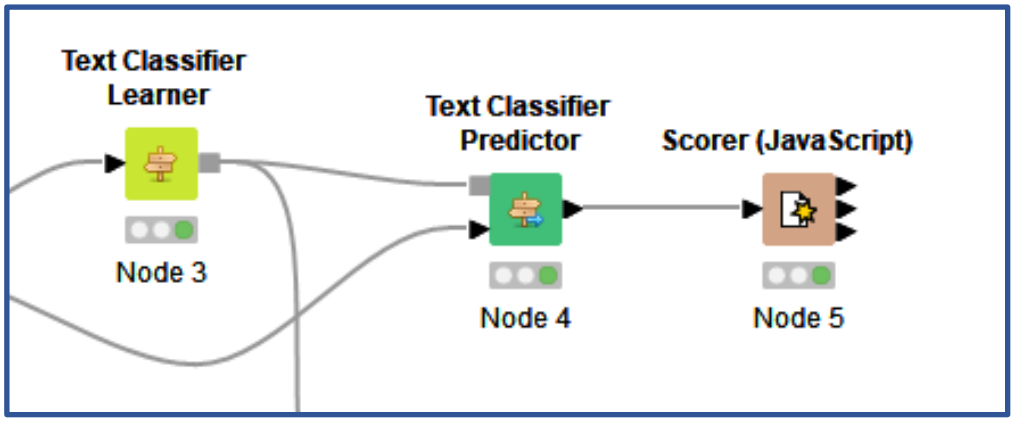
The second and last data preparation task is to split the dataset into training and test sets. To do that, we need a Partitioning node. As a general rule of thumb, we should use 80% of the data for training and the remaining 20% for testing. Drag a Partitioning node to the Workflow Editor. Then connect its input port to the output port of the excel Reader node. Double-click the Partitioning node and configure it as shown below:



We will use the option of stratified sampling with our sentiment *Label,* as shown in the above screenshot. Execute the Partitioning node when you have finished configuring it. After execution, the upper output port of the Partitioning node will provide the training set, while the lower output port will provide the test set. After executing, you should get something like the following (maybe without any further connections from the partitioning node).



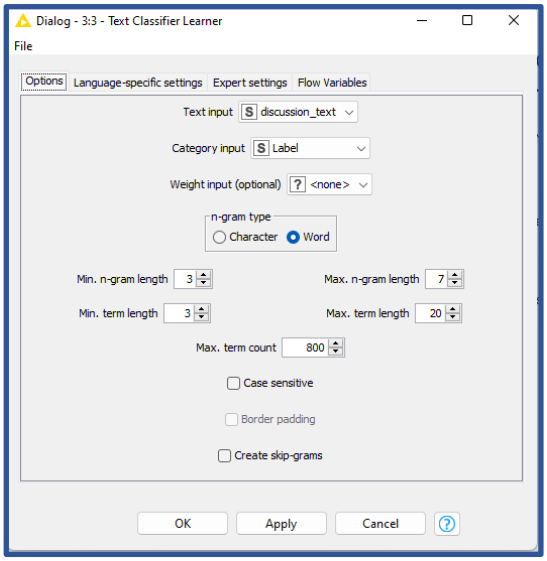
2.3 Build Sentiment Model



We will configure and train a *Text Classifier Learner, Text Classifier Predictor,* and *Scorer (Java Script)* nodes in this step.

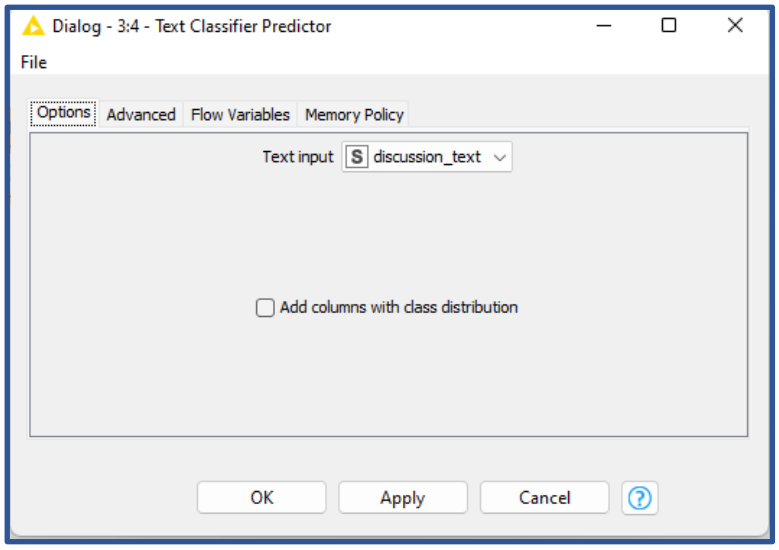
2.3.1 Text Classifier Learner node

Drag a *Text Classifier Learner* node to the Workflow Editor. Then connect its input port to the upper output port of the Partitioning node so that it will receive 80% of the data for training and building the sentiment model. Double-click the node to open the configuration options as shown below. Select the *discussion_text* as text input field and *Label* as category input, n-gram type as *Word* as shown below. Click OK to save the settings and close the node.



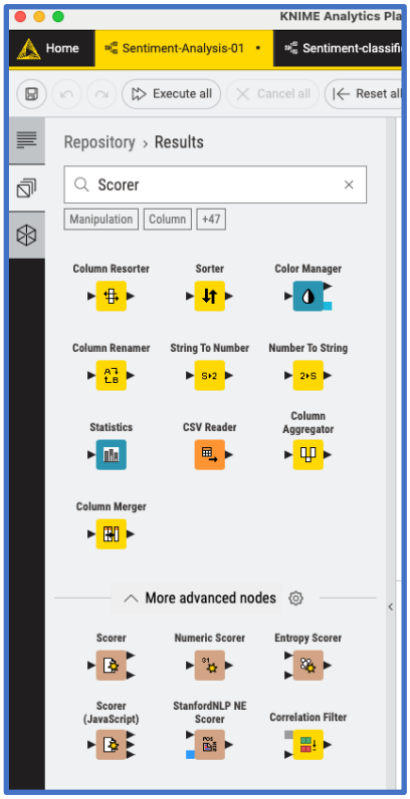
2.3.2 Text Classifier Predictor node

Connect the output from the *Text Classifier Learner* node to the input of the *Text Classifier Predictor* node, as shown in the above figure (figure in 2.3). Accept the default settings for the *Predictor* node as shown in the following figure, and you don't need to change anything if the *discussion_text* is selected as Text input.



2.3.3 Scorer node

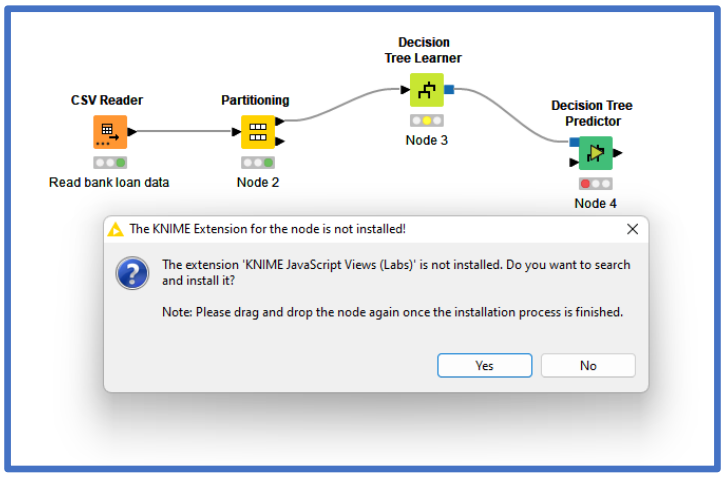
Search for the Scorer node in the node repository as shown in the following figure.



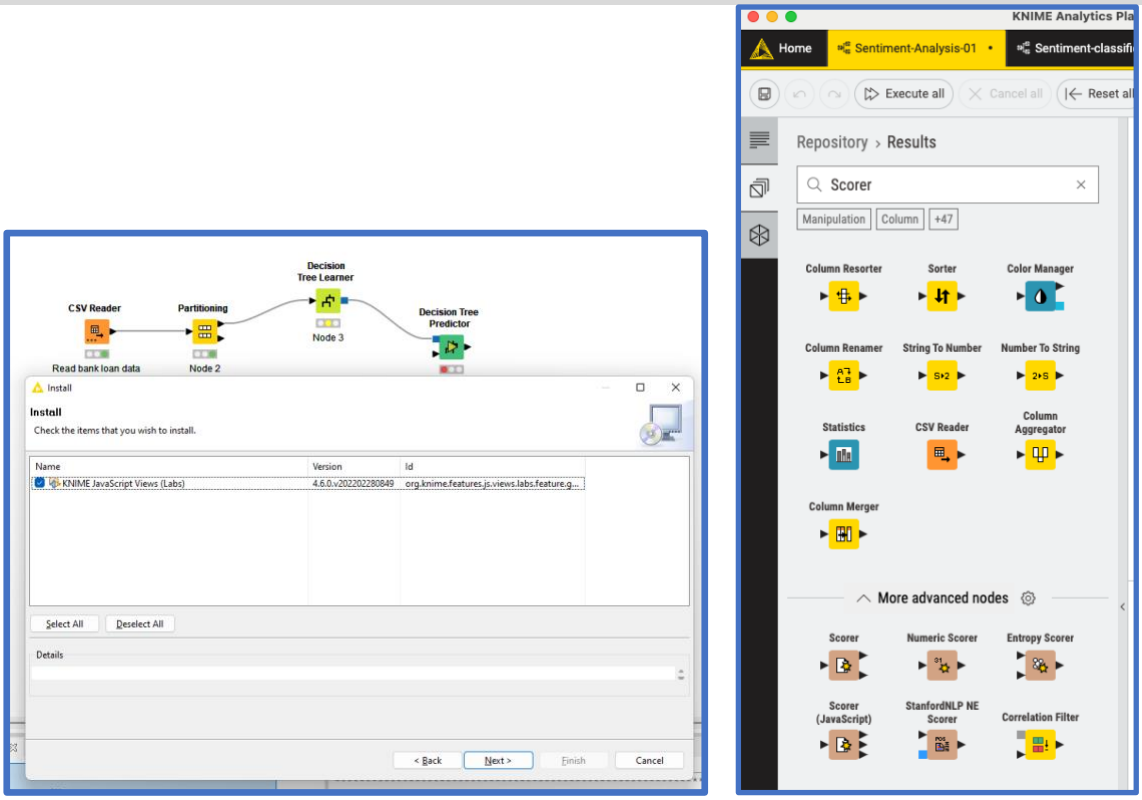
If you don't find the *Scorer (JavaScript)* node in the node repository, then follow the highlighted lines to add the *Scorer (JavaScript)* extension to KNIME.

Installing JavaScript scorer: JavaScript Scorer node is not installed in the KNIME platform by default. In order to use the node, first, you need to add/install the extension.

The best and easy way to install it is to go to the Scorer (JavaScript) webpage (https://hub.knime.com/knime/extensions/org.knime.features.js.views.labs/latest/org.knime.js.base.node.scorer.ScorerNodeFactory) and drag the  icon onto the KNIME workflow and it will automatically open a KNIME extension popup as shown below.

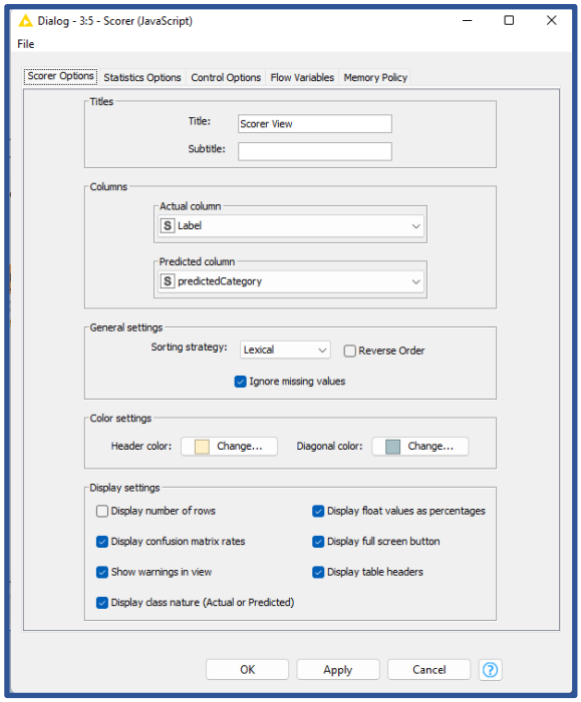


Press Yes to install the extension following the instructions in the below popup. After installing the extension, it will appear in the node repository when you search for it, as shown below.

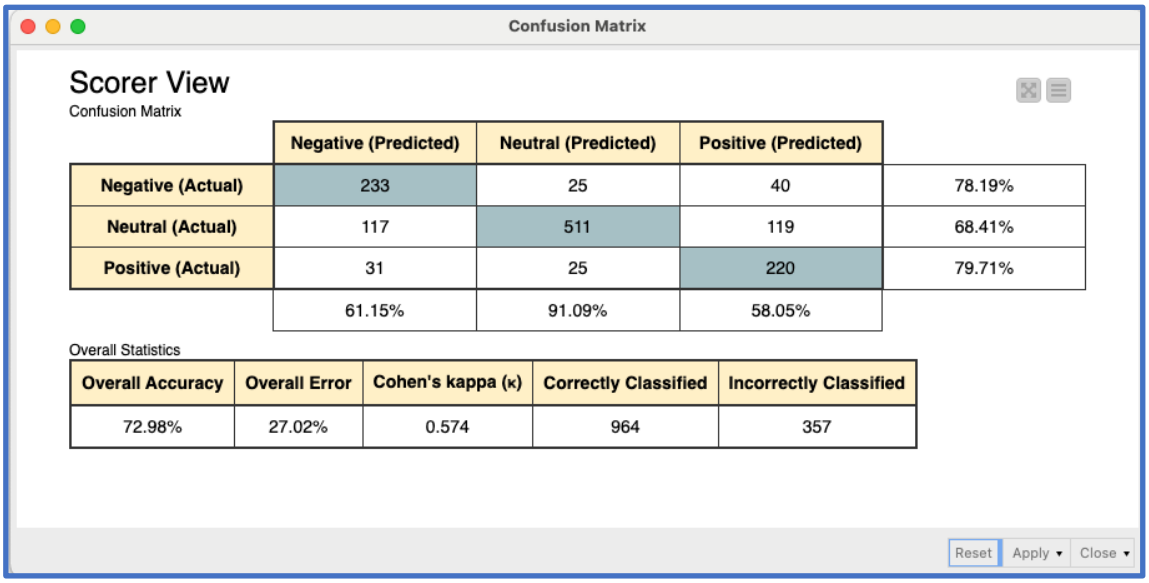


After installing the Scorer node, we can find it in the node repository.

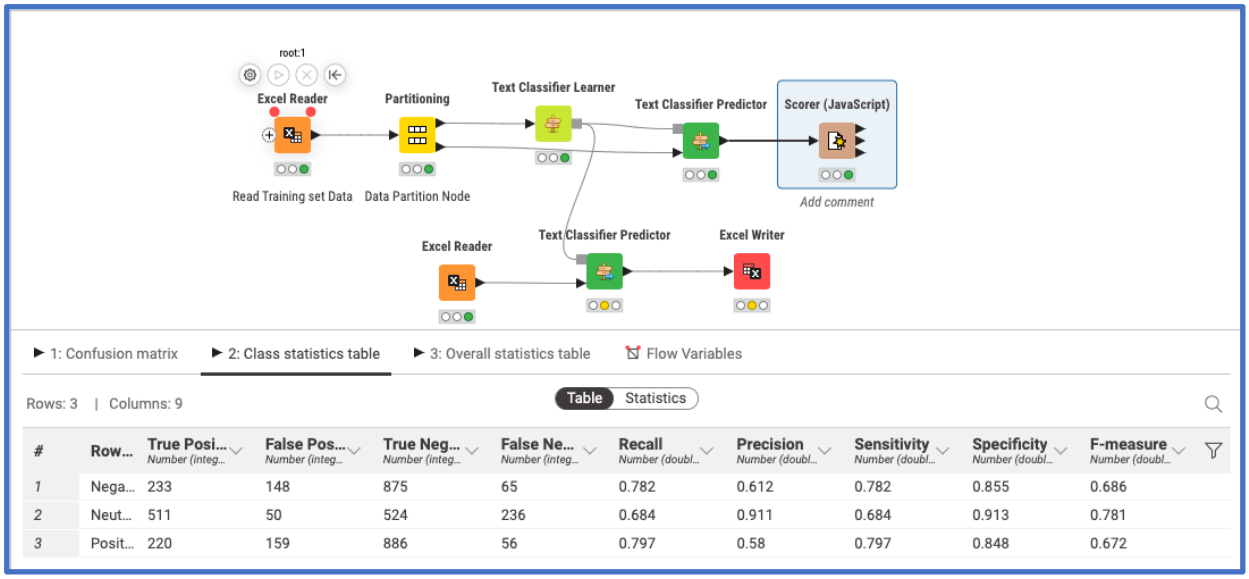
The scorer node calculates the statistics such as precision, recall, specificity, F-measure, and other measures. Search and drag the *Scorer* node on the workflow editor and connect it to the output of the *Text Classifier predictor* node, as shown in the workflow diagram at the beginning of this step (2.3 Build Sentiment Model). Choose the column names of the actual column (*Label*) and predicted column (*predictedCategory*) as shown in the figure of the following screenshot.



When you right-click the scorer node and select *Execute and Open Views* option to run all the nodes, then you will see the results in the following figure.



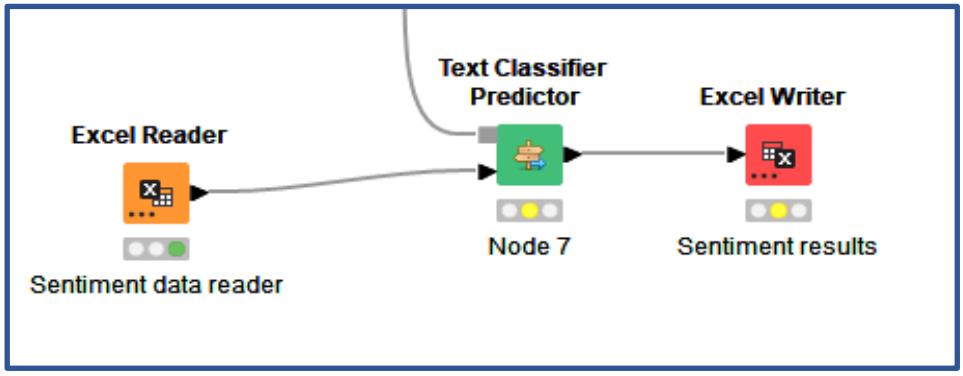
If you don't see this dialog, then you can right-click the node and select the *class statistics table* to view the following screenshot. As shown in the figure, you can see values for precision, recall, and F-measure.



#	Row...	True Posi...	False Pos...	True Neg...	False Ne...	Recall	Precision	Sensitivity	Specificity	F-measure
		Number (integ...	Number (integ...	Number (integ...	Number (integ...	Number (doubl...	Number (doubl...	Number (doubl...	Number (doubl...	Number (doubl...
1	Nega...	233	148	875	65	0.782	0.612	0.782	0.855	0.686
2	Neut...	511	50	524	236	0.684	0.911	0.684	0.913	0.781
3	Posit...	220	159	886	56	0.797	0.58	0.797	0.848	0.672

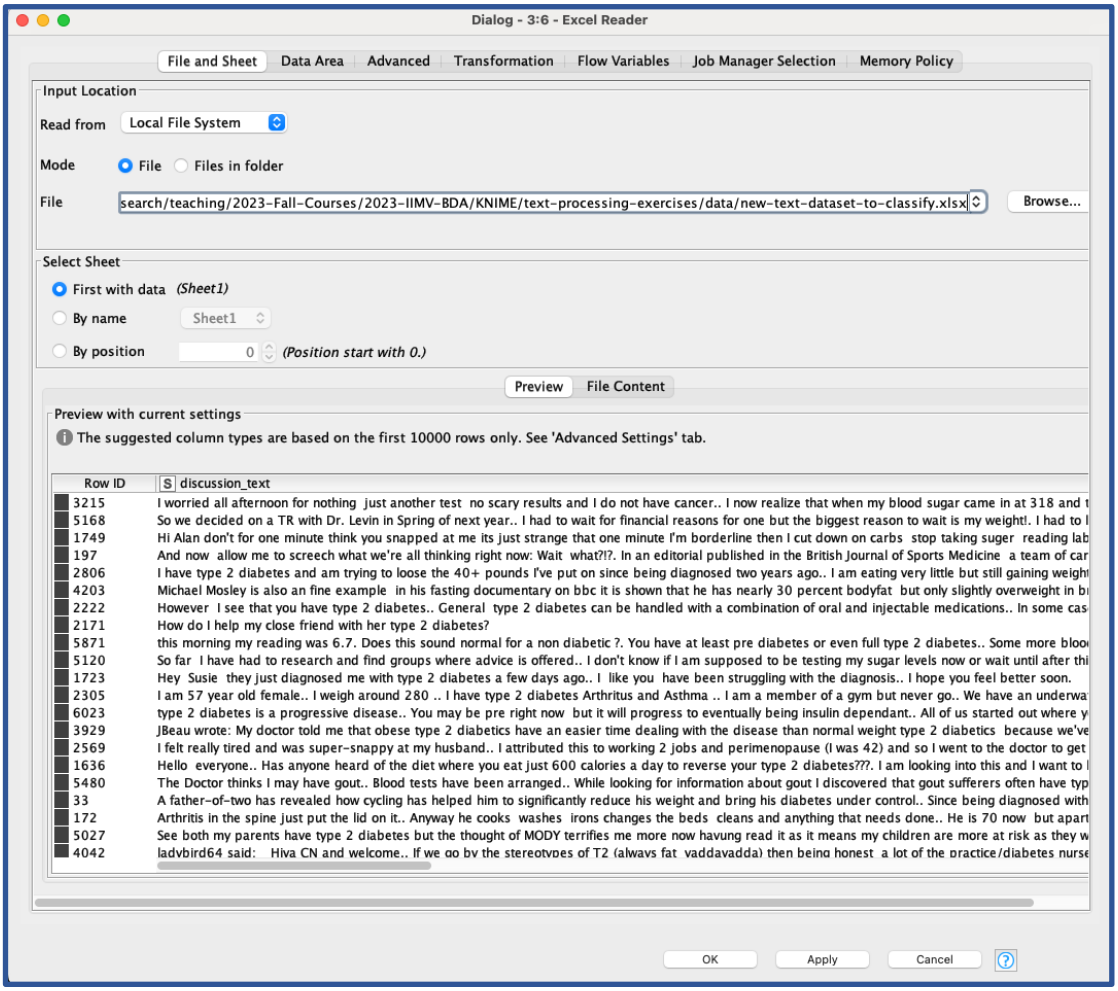
2.4 Make predictions for new discussions

Finally, when once the Sentiment model results are satisfactory, we can use it to predict the sentiment of a given text. This step contains three nodes, as shown in the following screenshot. Please search for the nodes Excel Reader, Text Classifier Predictor, and Excel writer and connect them as shown in the following screenshot of the KNIME workflow.



2.4.1 Excel Reader

In this step, we will use *new-text-dataset-to-classify.xlsx,* which contains new textual data from discussion forums on which we want to make sentiment predictions. Follow the same procedure as described in the first step for reading data from the *discussion-forums-sentiment-training-set-data.xlsx* file from section 2.2.1. The data in *new-text-dataset-to-classify.xlsx* is in the same format as that *discussion-forums-sentiment-training-set-data.xlsx*, containing the input *discussion_text* column except for the *Label* column, which we will predict using the sentiment model we have trained previously. The configuration settings of the Excel reader are shown below.

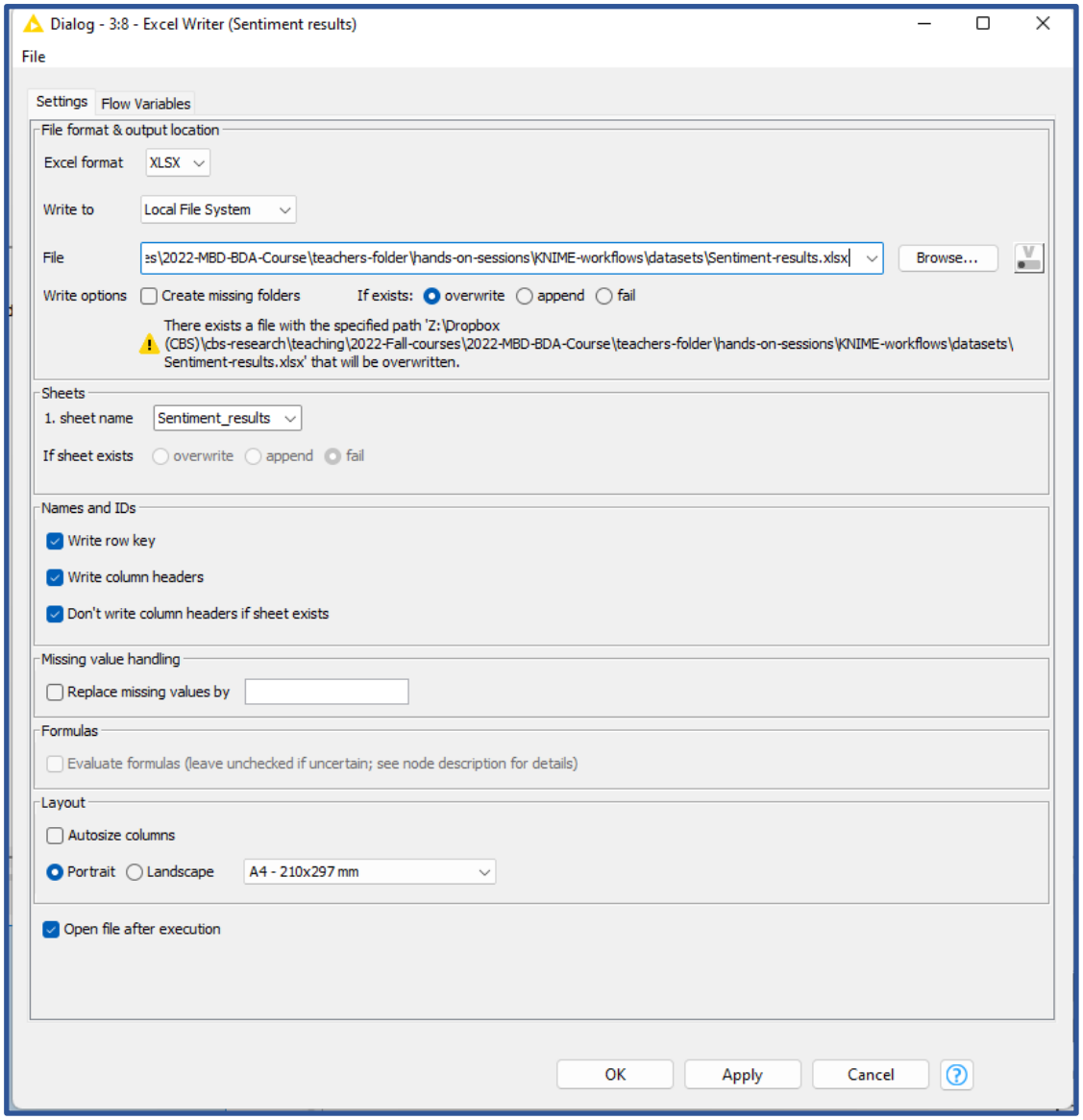


2.4.2 Text Classifier Predictor node

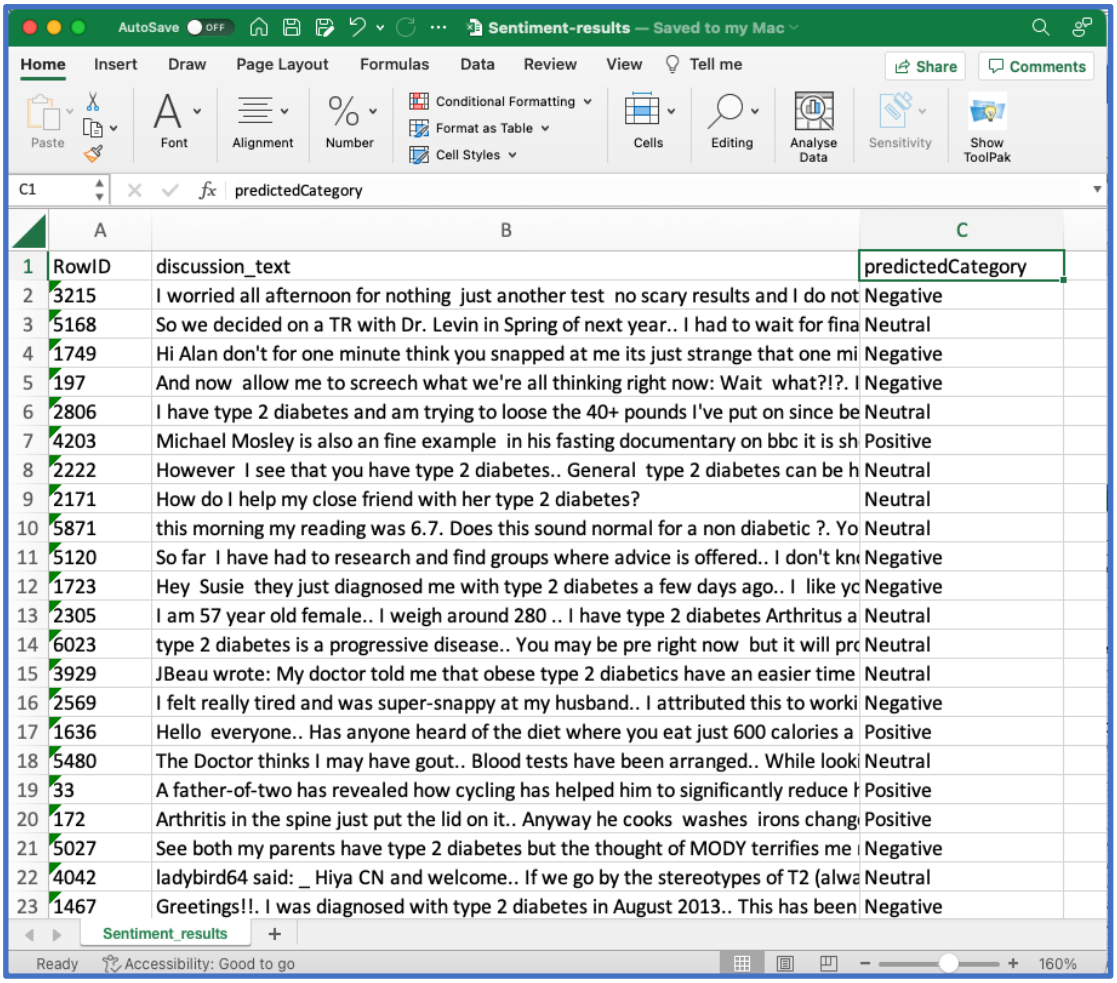
Connect the output from the *Text Classifier Learner* node to the input of the *Text Classifier Predictor* node, as shown in the above figure in 2.4. Accept the default settings for the *Text Classifier Predictor* node as previously done in the training step.

2.4.3 Excel writer node

Finally, we would like to save the sentiment predictions made on the new text data into a file using excel writer. Double-click the Excel writer node to open its configuration dialog. Choose the configuration options as shown in the following figure. Choose an appropriate path for the file using the browse button, specify the file name and select the overwrite option if it exists. You can also choose the write row key and write column headers options. Also, choose *open file after execution* to open the Excel file after execution. Finally, click ok to save the configuration options and then execute the node to save the predictions of bank loan results into an excel sheet.



Once the node completes its execution, you can see the new predictions of sentiments in the *predictedCategory* as shown in the following screenshot.



AutoSave OFF — Sentiment-results — Saved to my Mac

Home Insert Draw Page Layout Formulas Data Review View Tell me — Share — Comments

Paste | Font | Alignment | Number | Conditional Formatting | Format as Table | Cell Styles | Cells | Editing | Analyse Data | Sensitivity | Show ToolPak

C1 — predictedCategory

	A	B	C
1	RowID	discussion_text	predictedCategory
2	3215	I worried all afternoon for nothing just another test no scary results and I do not	Negative
3	5168	So we decided on a TR with Dr. Levin in Spring of next year.. I had to wait for fina	Neutral
4	1749	Hi Alan don't for one minute think you snapped at me its just strange that one mi	Negative
5	197	And now allow me to screech what we're all thinking right now: Wait what?!?. I	Negative
6	2806	I have type 2 diabetes and am trying to loose the 40+ pounds I've put on since be	Neutral
7	4203	Michael Mosley is also an fine example in his fasting documentary on bbc it is sh	Positive
8	2222	However I see that you have type 2 diabetes.. General type 2 diabetes can be h	Neutral
9	2171	How do I help my close friend with her type 2 diabetes?	Neutral
10	5871	this morning my reading was 6.7. Does this sound normal for a non diabetic ?. Yo	Neutral
11	5120	So far I have had to research and find groups where advice is offered.. I don't kn	Negative
12	1723	Hey Susie they just diagnosed me with type 2 diabetes a few days ago.. I like yc	Negative
13	2305	I am 57 year old female.. I weigh around 280 .. I have type 2 diabetes Arthritus a	Neutral
14	6023	type 2 diabetes is a progressive disease.. You may be pre right now but it will pr	Neutral
15	3929	JBeau wrote: My doctor told me that obese type 2 diabetics have an easier time	Neutral
16	2569	I felt really tired and was super-snappy at my husband.. I attributed this to worki	Negative
17	1636	Hello everyone.. Has anyone heard of the diet where you eat just 600 calories a	Positive
18	5480	The Doctor thinks I may have gout.. Blood tests have been arranged.. While looki	Neutral
19	33	A father-of-two has revealed how cycling has helped him to significantly reduce h	Positive
20	172	Arthritis in the spine just put the lid on it.. Anyway he cooks washes irons chang	Positive
21	5027	See both my parents have type 2 diabetes but the thought of MODY terrifies me	Negative
22	4042	ladybird64 said: _ Hiya CN and welcome.. If we go by the stereotypes of T2 (alwa	Neutral
23	1467	Greetings!!. I was diagnosed with type 2 diabetes in August 2013.. This has been	Negative

Sentiment_results

Ready — Accessibility: Good to go — 160%