ABBYY® Smart Classifier 2.7

User Guide

Table of Contents

Table of Contents

Introducing ABBYY Smart Classifier

Document classification is the process of putting documents into categories based on their features and properties.

Documents can be classified manually or automatically.

Manual classification of large numbers of documents has the following problems associated with it:

- It is labor-intensive.
- It is expensive because it requires a lot of classification specialists.
- It is slow and cannot be used in projects where time is of the essence.
- Classification quality deteriorates when classification specialists have to work to tight deadlines.

With ABBYY Smart Classifier, you can classify documents automatically, avoiding most of the problems associated with manual document classification.

ABBYY Smart Classifier uses ABBYY Compreno text analysis and understanding technology to categorize documents. It can be easily integrated into document management systems, knowledge bases, and other systems that work with structured and unstructured data.

The benefits of using ABBYY Smart Classifier

ABBYY Smart Classifier can be used to classify texts in many different languages, assigning documents to their appropriate categories as instructed by the user. When classifying documents, ABBYY Smart Classifier looks for certain features which are characteristic of each category of documents. In the case of Russian, English, and German texts, ABBYY Compreno semantic analysis technology can be applied in order to classify documents based on the meaning of analyzed texts.

ABBYY Smart Classifier offers an intuitive graphical user interface and does not require any special skills from the end user.

ABBYY Smart Classifier can be used in a wide range of tasks that require the processing of vast amounts of unstructured information, such as routing incoming documents to their appropriate departments, forwarding letters and e-mail messages to their intended addressees, determining how long different kinds of documents should be stored in a company's information system, and many more. ABBYY Smart Classifier automates a lot of classification chores, simplifying a lot of business processes and enabling employees to navigate their way through enormous masses of information. By automating their document classification routines, companies will radically speed up document processing and avoid human errors that are almost inevitable when large volumes of data are classified manually.

Usage scenarios

ABBYY Smart Classifier can be effectively used in the following scenarios:

- **Handling requests from members of the public**

  The procedures for handling of citizens' complaints, requests, petitions, and appeals by government bodies are regulated by law. With ABBYY Smart Classifier, government institutions can automate the classification of incoming documents into predefined topics, reducing the number of classification errors and response times.

- **Analyzing technical support requests**

  Technical support requests are an important source of customer feedback data. Support engineers spend a lot of their time on grouping incoming requests into categories. ABBYY Smart Classifier allows you to automate this work, enabling your support engineers to spend more time on solving customers' issues.

- **Selecting document storage policies**

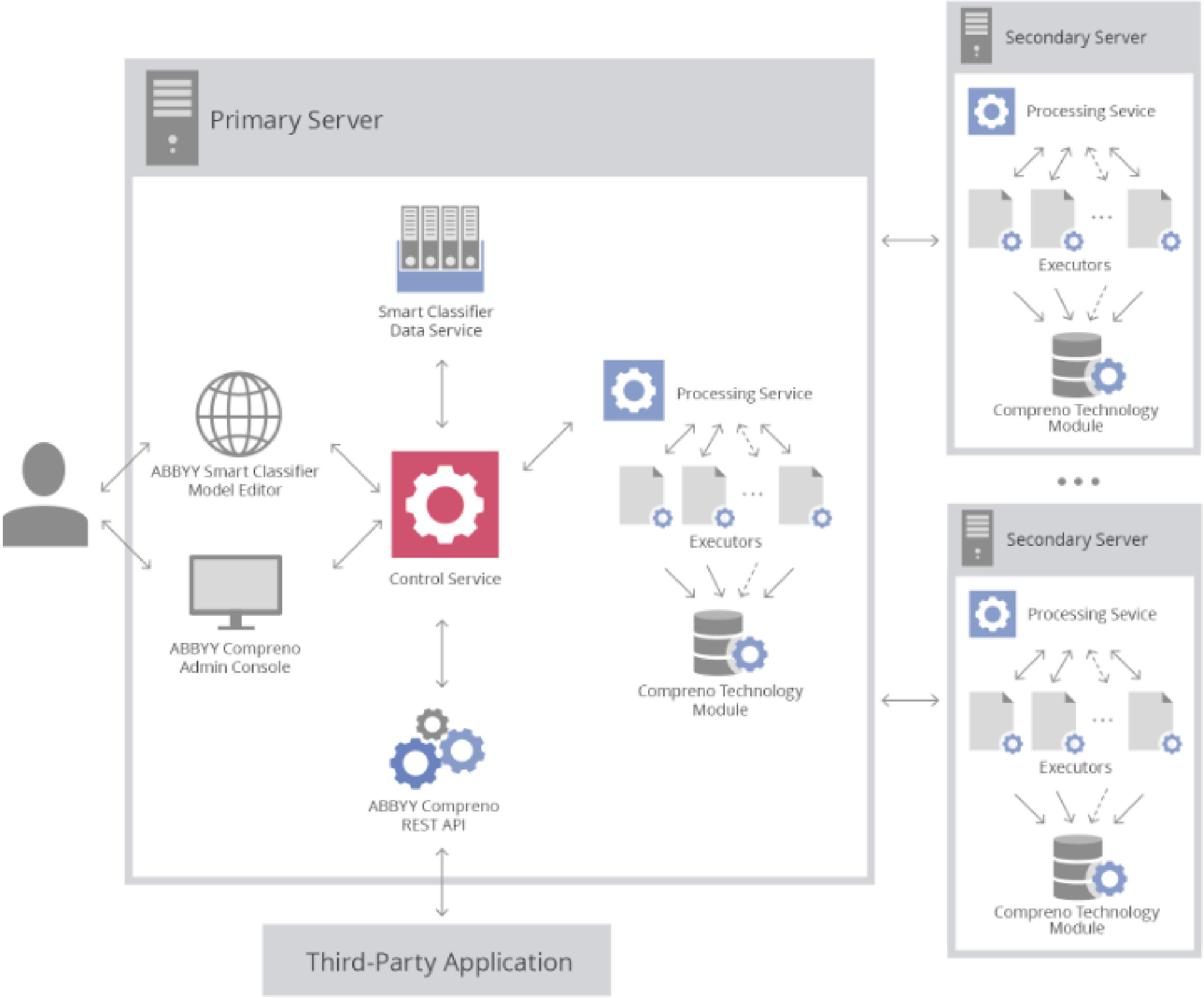
  Companies that receive huge amounts of documents on a daily basis can use ABBYY Smart Classifier to classify incoming and archived documents for storage purposes, so that they can implement different storage policies for different types of documents.

- **Assigning attributes to documents**

  The amount of information available to users increases exponentially, and more sophisticated tools are required to find and retrieve relevant documents. ABBYY Smart Classifier can be used to add attributes to documents in your collections for faster and more reliable search experience.

ABBYY Smart Classifier architecture

The architecture of ABBYY Smart Classifier is shown in the figure below.



The function and purpose of each ABBYY Smart Classifier component is described below.

- The **Control Service** distributes the workload among the available Processing Services and interacts with all the other ABBYY Smart Classifier components.
- The **Processing Service** processes tasks received from the Control Service.
- The **Compreno Technology Module** provides text classification algorithms.
- The **ABBYY Compreno Admin Console** is used for administering ABBYY Smart Classifier.
- The **ABBYY Compreno REST API** enables integration of ABBYY's classification technologies into third-party systems.
- The **Smart Classifier Data Service** enables the use of classification models.
- The **ABBYY Smart Classifier Model Editor** is where you create, train, and deploy classification models.

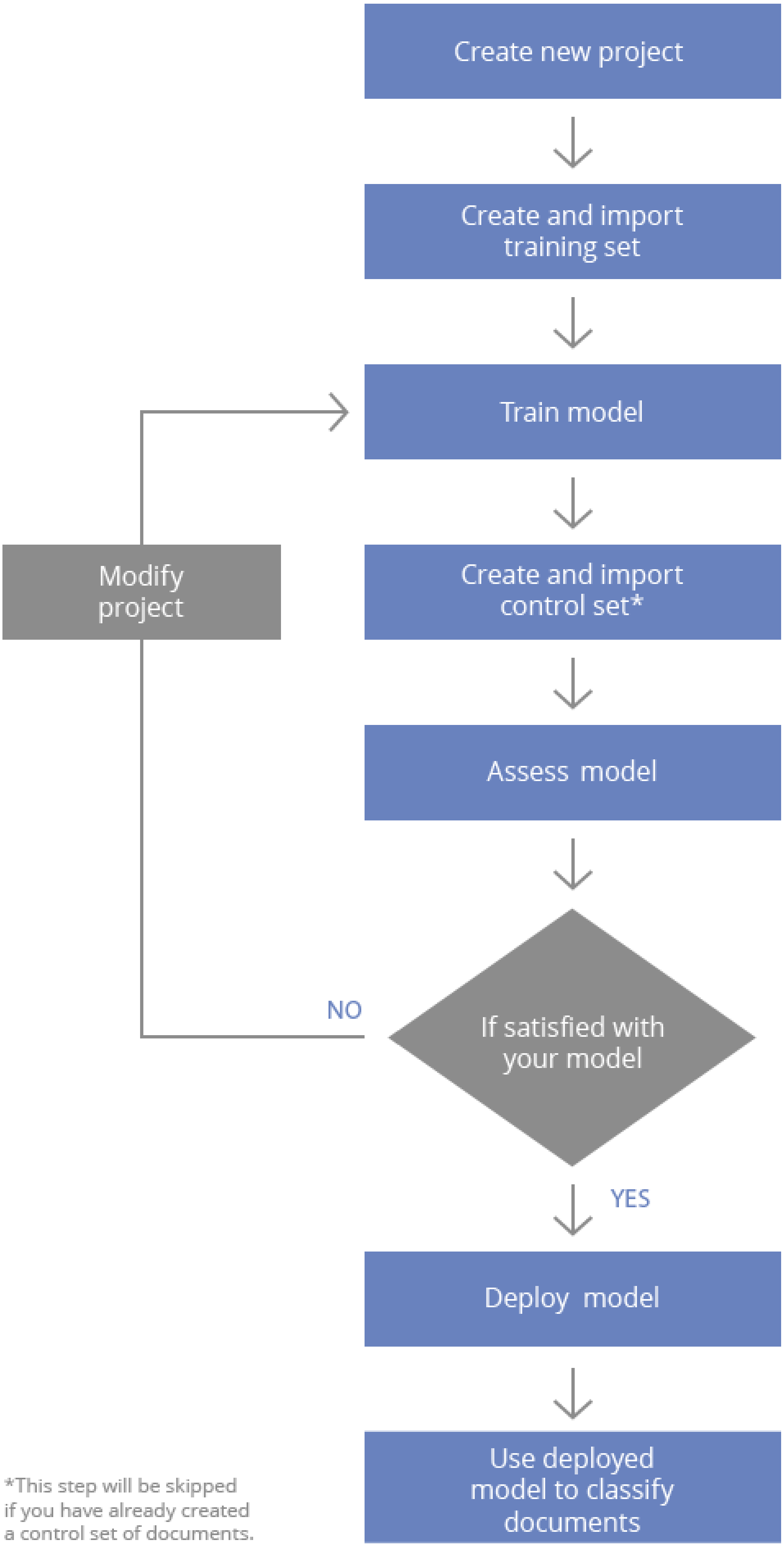
About Document Classification

ABBYY Smart Classifier uses machine learning methods for automatic document classification, dynamically adapting to any changes in the composition and content of document collections. Here is how it works.

Suppose you have a set of documents which have been manually categorized by experts, say, newswire articles categorized as SPORT, POLITICS, BUSINESS, etc. ABBYY Smart Classifier can use these documents to train itself to classify any other newswire articles that are fed to it. A set of manually categorized documents is termed a training set. ABBYY Smart Classifier uses a training set to create a classification model, which enables it to classify any other documents that have features similar to those of the documents in the training set.

The life cycle of a classification model

The life cycle of a classification model is shown in the figure below. All of the steps in this flow chart are performed in the ABBYY Smart Classifier Model Editor.


Create new project
Create and import training set
Train model
Modify project
Create and import control set*
Assess model
NO
If satisfied with your model
YES
Deploy model
Use deployed model to classify documents

*This step will be skipped
if you have already created
a control set of documents.

1. Create a new project

Create a new project and specify the settings for a classification model.

2. Create and import a training set of documents

Create a training set of documents, where all documents are assigned their appropriate categories. These documents will be used for training your model. Next, import the training set into your project.

3. Train your model

When you train your model, ABBYY Smart Classifier automatically determines:

- features specific to each category
- an algorithm that provides the best quality of classification

4. Create and import a control set of documents

Create a control set of documents, where all documents are assigned their appropriate categories. The documents in the control set must not be the same as in the training set. These documents will be used for assessing the classification power of your model. Next, import the control set into your project.

5. Assess your model

Use your trained model to classify the documents in the control set. The classification results you obtain on the control set are a reliable indicator of the quality of classification you may obtain on any other set of documents. If you are satisfied with the quality of classification, deploy the model. Otherwise, modify the project until you are satisfied with the results.

6. Modify your project*

***This is an optional step to be performed if you are not satisfied with your model.**

To improve your classification model, modify the project settings or the training set and update your model.

7. Deploy your model

Once deployed, your model will be available for document classification by means of the ABBYY Compreno REST API or on the **Document Classification** page of the ABBYY Smart Classifier Model Editor.

Getting Started

The purpose of this section is to help you get started with the ABBYY Smart Classifier Model Editor using a sample collection of documents. You can find the sample documents in the following folder:

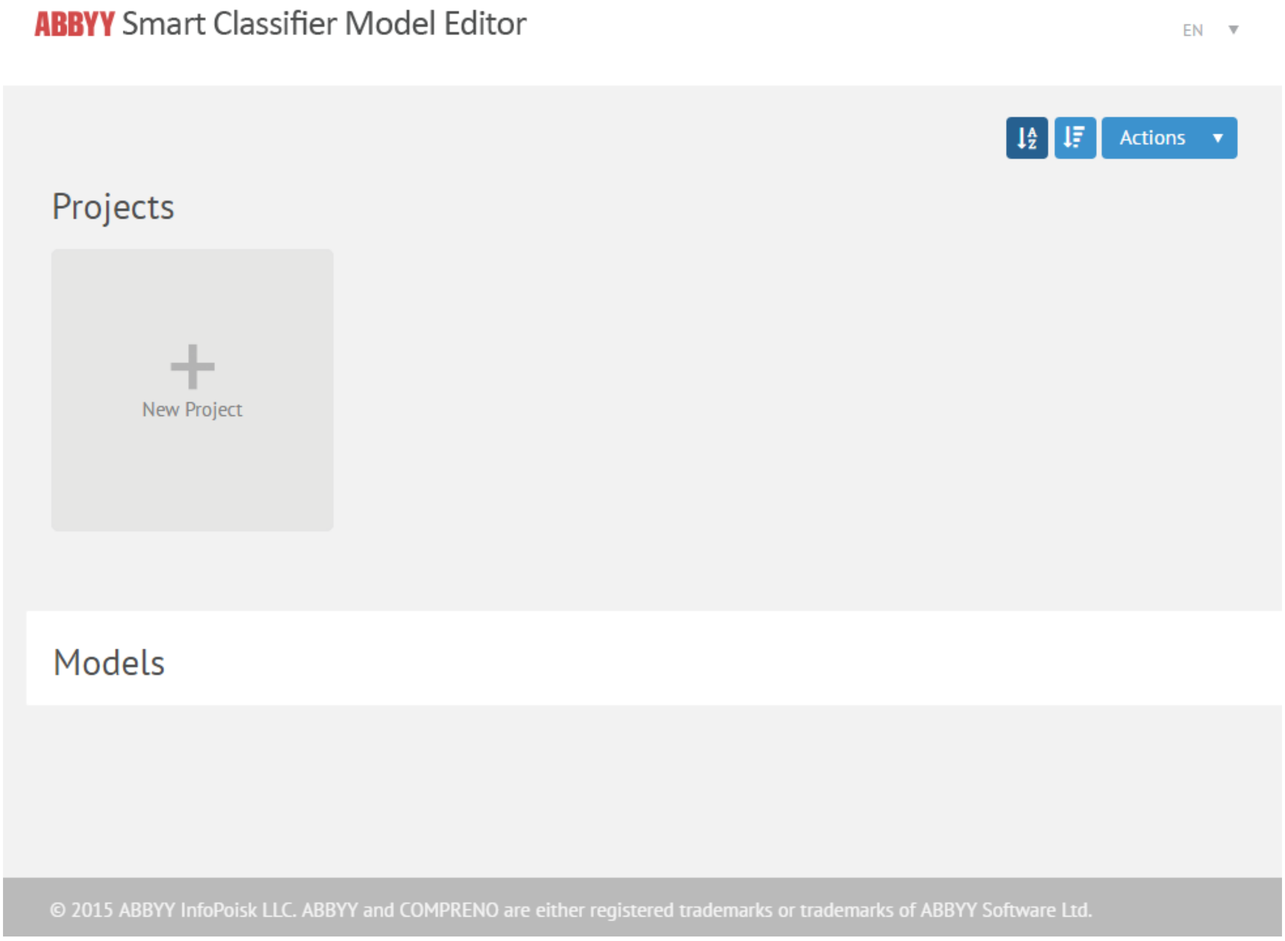
%PUBLIC%\ABBYY\Compreno Products\2.7\Code Samples\SmartClassifierSampleApplication\SampleSets

To be able to classify documents, you need to create and deploy a classification model. To create a classification model, you must create a new project first.

Creating a project

To create a project, complete these steps:

1. Start the ABBYY Smart Classifier Model Editor by clicking **Start > ABBYY Compreno Products > ABBYY Smart Classifier Model Editor**.


ABBYY Smart Classifier Model Editor
EN ▼
Actions ▼
Projects
New Project
Models
© 2015 ABBYY InfoPoisk LLC. ABBYY and COMPRENO are either registered trademarks or trademarks of ABBYY Software Ltd.

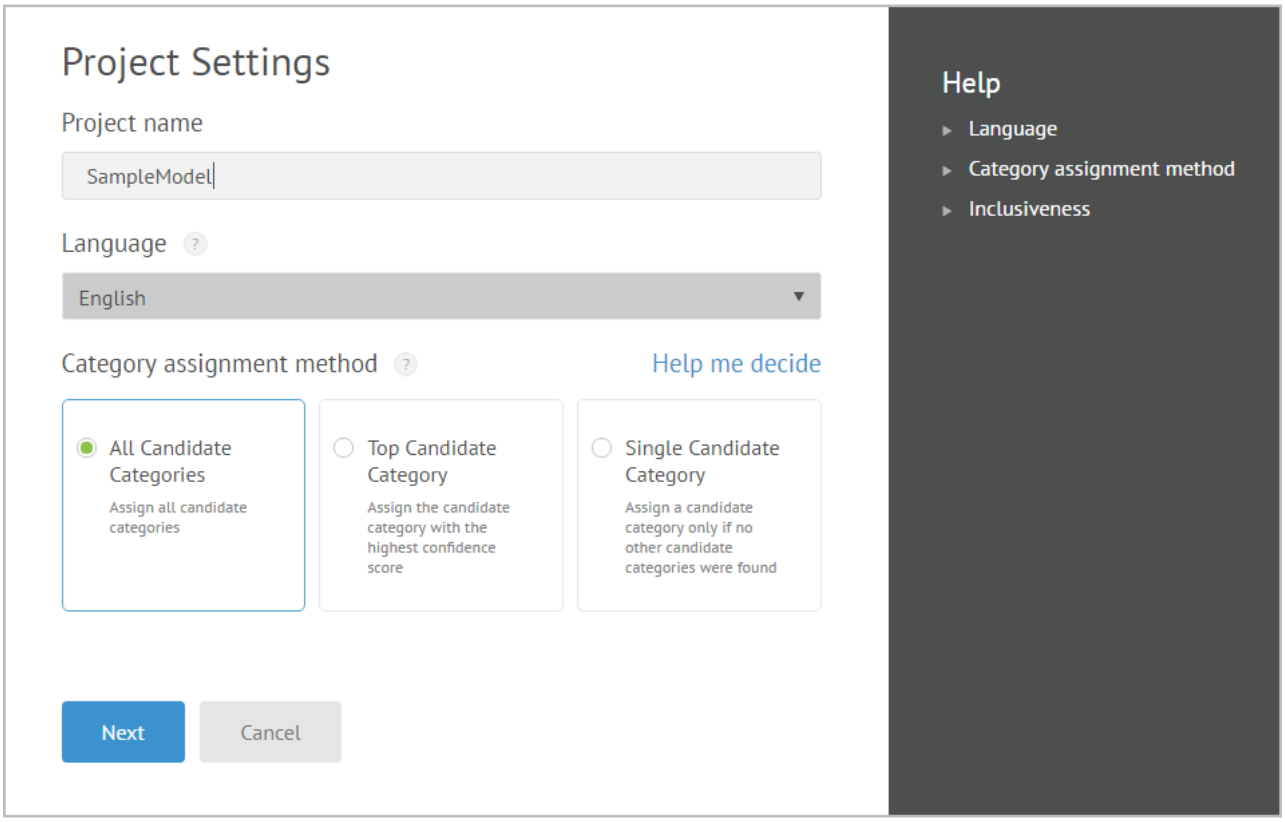
2. Click the **RU/EN** icon in the top right corner to choose Russian or English as the GUI language of the editor.
3. Click the **New Project** tile.
4. On the screen that opens, set up your project (refer to the information panel for help on the project settings).

   Type this name for your new project: *SampleModel*.

   Select English as the language of your documents.

   Select a category assignment method.

   Click the **Next** button.



Project Settings

Project name

SampleModel

Language ?

English ▾

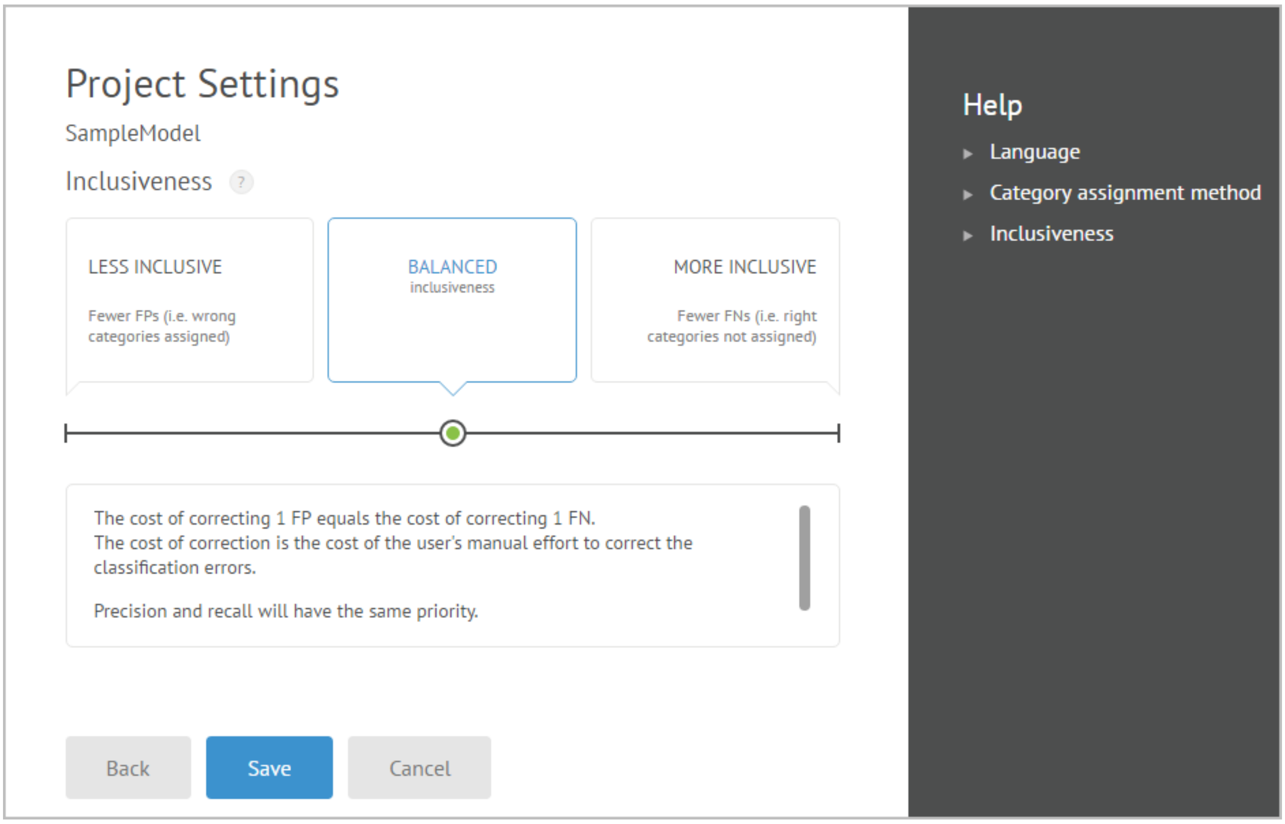
Category assignment method ? Help me decide

● All Candidate Categories
Assign all candidate categories

○ Top Candidate Category
Assign the candidate category with the highest confidence score

○ Single Candidate Category
Assign a candidate category only if no other candidate categories were found

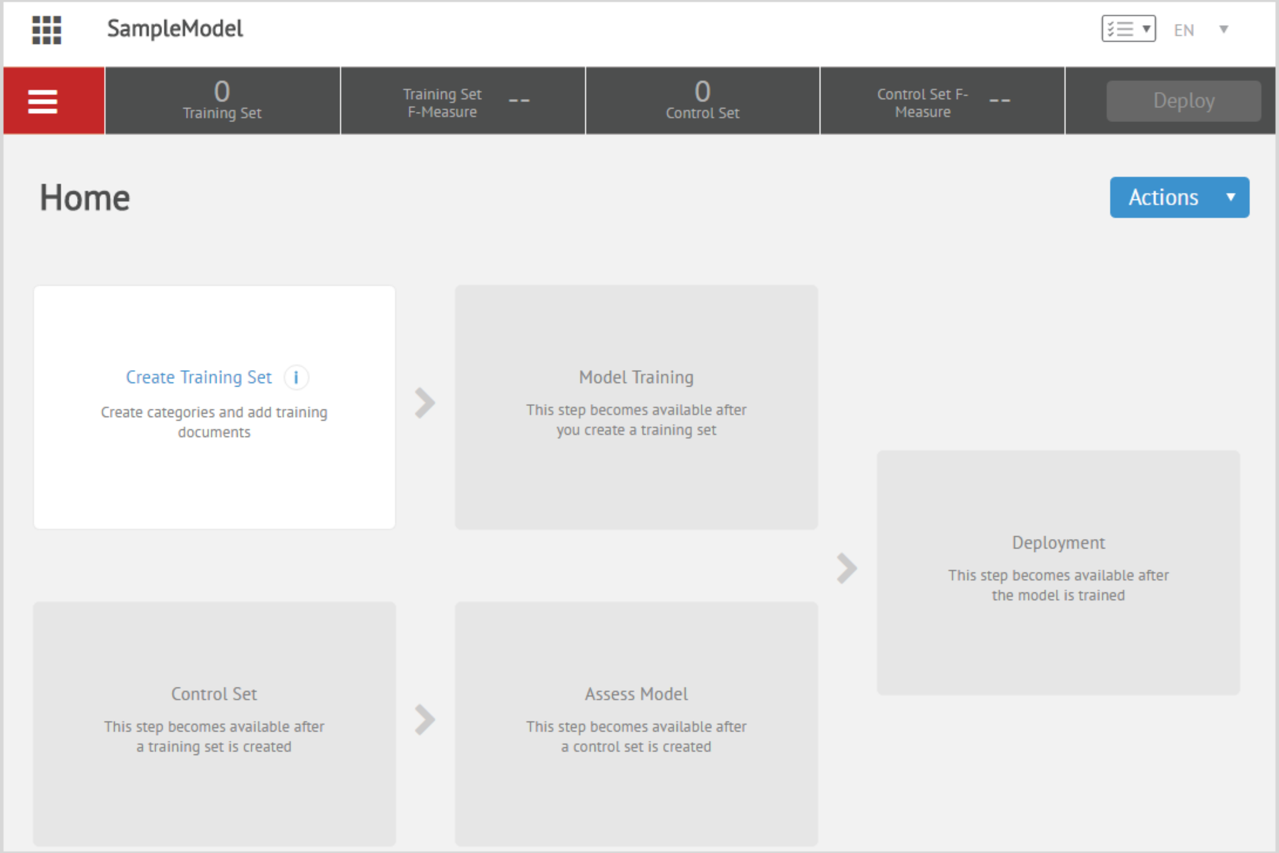
5. On the next screen, move the **Inclusiveness** slider to the left or to the right, depending on which is more important in your project — precision or recall. Then click the **Save** button to save your settings.



The cost of correcting 1 FP equals the cost of correcting 1 FN.
The cost of correction is the cost of the user's manual effort to correct the classification errors.

Precision and recall will have the same priority.

The home page of your newly created project will open, where you can see the steps to be performed in order to obtain a working classification model. Click the i icon on each tile to see a detailed description of the respective step.



SampleModel
EN
0 Training Set
Training Set F-Measure --
0 Control Set
Control Set F-Measure --
Deploy
Home
Actions
Create Training Set i
Create categories and add training documents
Model Training
This step becomes available after you create a training set
Deployment
This step becomes available after the model is trained
Control Set
This step becomes available after a training set is created
Assess Model
This step becomes available after a control set is created

Training, assessing and deploying your model

Step 1. Create a training set of documents

To train your model, you need to import a training set of documents. To create and import a training set:

1. Click the **Create Training Set** tile.
2. In the dialog box that opens, click the **Import training set** tile and follow the instructions that appear on the screen.

   Select the *Training set.zip* training set that is provided with the ABBYY Compreno REST API sample code. This training set can be found in:

   %PUBLIC%\ABBYY\Compreno Products\2.7\Code Samples\SmartClassifierSampleApplication\SampleSets

   After you import the training set, training will start automatically.

Step 2. Train your model

As a result, a classification model will be created. Training times depend on the number of categories and documents in the training set.

Important! We recommend assessing your model on a control set of documents (steps 3 and 4 below) in order to see how your model performs on documents other than those included in the training set.

Step 3. Create a control set of documents

A control set contains documents to which their appropriate categories have been assigned by the user and which are used to assess the classification power of the model. The documents in the control set are different from those used for training the model.

To create and import a control set:

1. Click the **Control Set** tile.
2. In the dialog box that opens, import the archive that contains the documents to be included in the control set.

   Select the *Control set.zip* control set that is provided with the ABBYY Compreno REST API sample code. This control set can be found in:

%PUBLIC%\ABBYY\Compreno Products\2.7\Code Samples\SmartClassifierSampleApplication\SampleSets

After you import the control set, the assessment will start automatically.

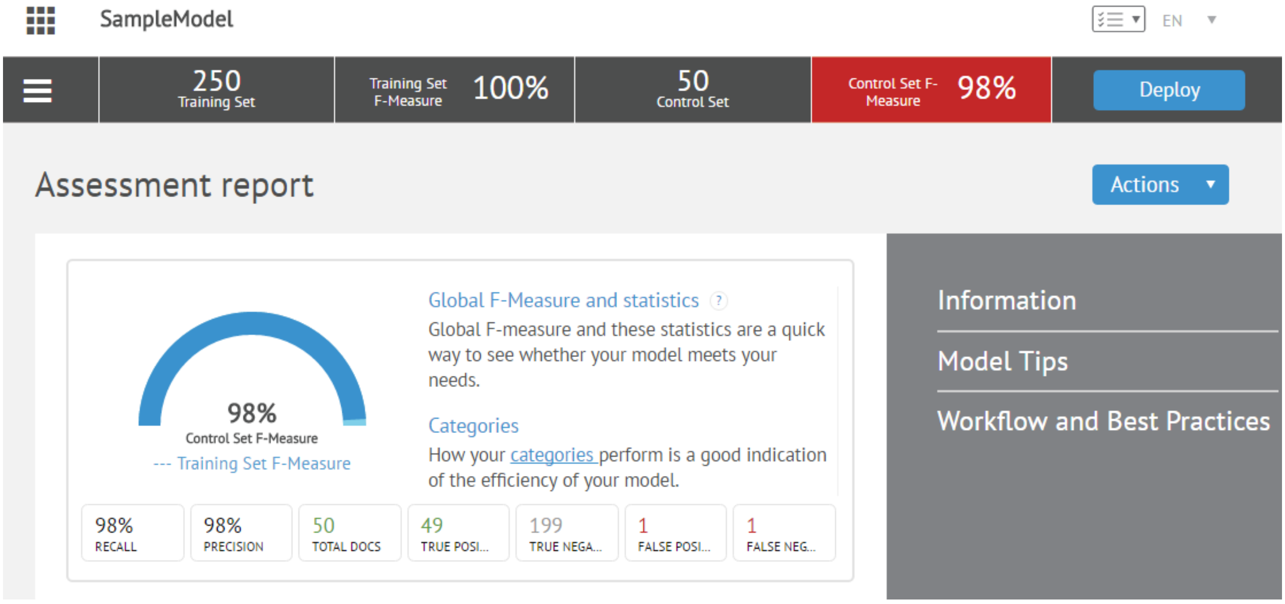
Step 4. Assess your model

To assess the classification power of your trained model, assess it using the control set of documents.

After you classify the documents in the control set, an assessment report is generated. The report contains the following statistics:

- F-measure
- Recall
- Precision
- The number of TPs, TNs, FPs, and FNs

These statistics are an objective assessment of your model's classification power, because the control set does not contain any documents used for training the model.



Global F-Measure and statistics

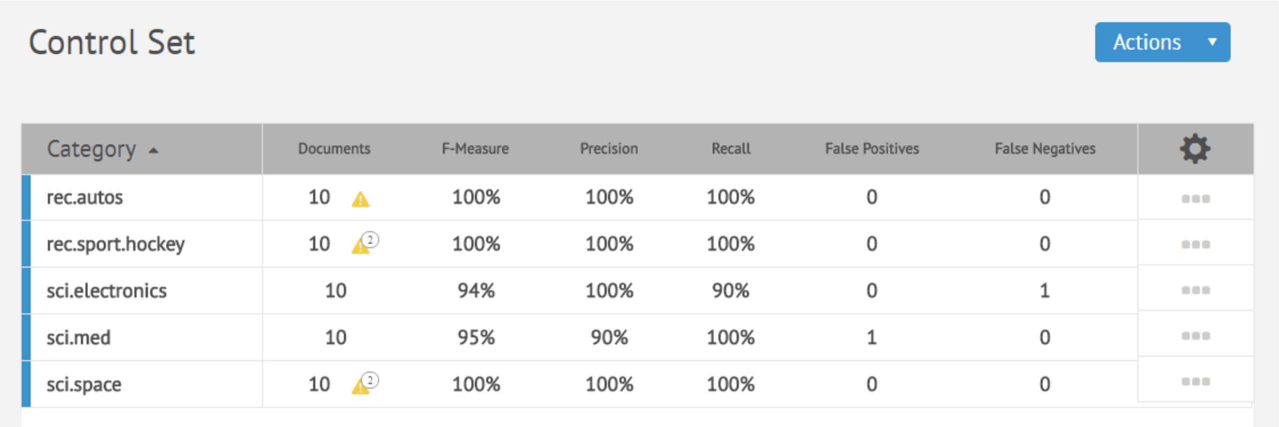
Global F-measure and these statistics are a quick way to see whether your model meets your needs.

Categories

How your categories perform is a good indication of the efficiency of your model.

98%	98%	50	49	199	1	1
RECALL	PRECISION	TOTAL DOCS	TRUE POSI...	TRUE NEGA...	FALSE POSI...	FALSE NEG...

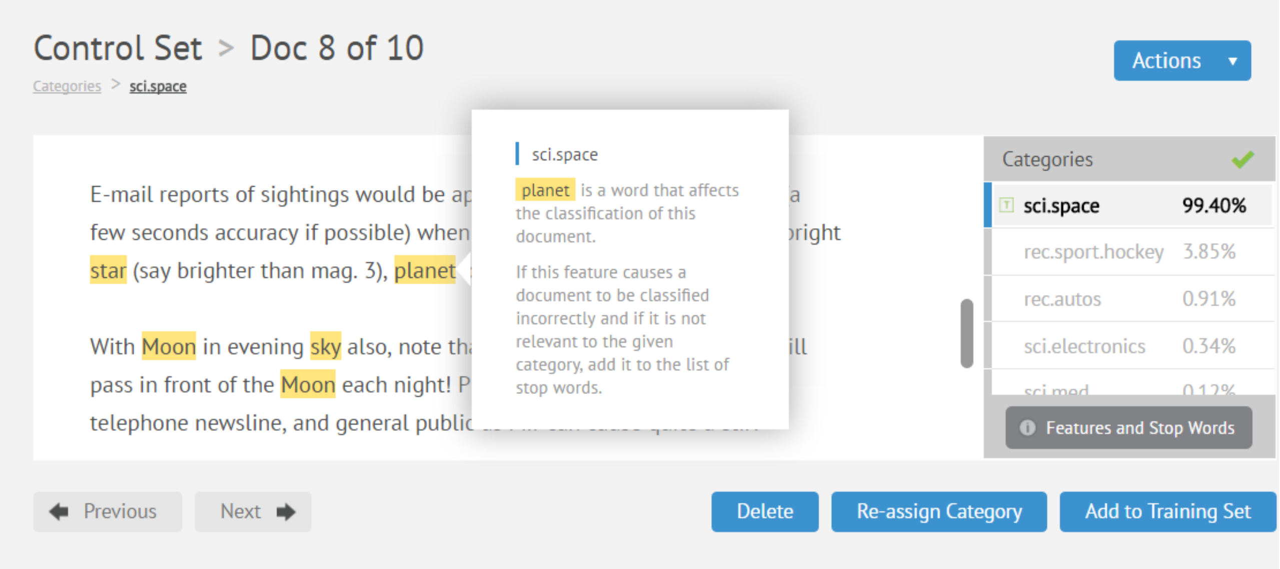
To see the statistics for each category, open the page with the list of document categories. You can use this information to see how the model performs for each category.



Control Set

Category ▲	Documents	F-Measure	Precision	Recall	False Positives	False Negatives	⚙
rec.autos	10 ⚠	100%	100%	100%	0	0	•••
rec.sport.hockey	10 ⚠	100%	100%	100%	0	0	•••
sci.electronics	10	94%	100%	90%	0	1	•••
sci.med	10	95%	90%	100%	1	0	•••
sci.space	10 ⚠	100%	100%	100%	0	0	•••

Open a category to see the documents to which a true category has been wrongly not assigned. Open a document to see the features that made the program place it in this particular category (highlighted in orange).



Control Set > Doc 8 of 10

Categories > sci.space

E-mail reports of sightings would be ap... a few seconds accuracy if possible) when ... bright star (say brighter than mag. 3), planet ...

With Moon in evening sky also, note th... ill pass in front of the Moon each night! P... telephone newsline, and general public ...

sci.space

planet is a word that affects the classification of this document.

If this feature causes a document to be classified incorrectly and if it is not relevant to the given category, add it to the list of stop words.

Categories ✓	
sci.space	99.40%
rec.sport.hockey	3.85%
rec.autos	0.91%
sci.electronics	0.34%
sci.med	0.12%

If you are satisfied with the trained model, deploy it. Otherwise, follow these recommendations to improve your model.

Step 5. Deploy your model

To deploy your model, click the **Deploy Model** tile.

Once deployed, the model will become available for document classification via the ABBYY Compreno REST API and on the **Document Classification** page of the ABBYY Smart Classifier Model Editor.

Classifying documents

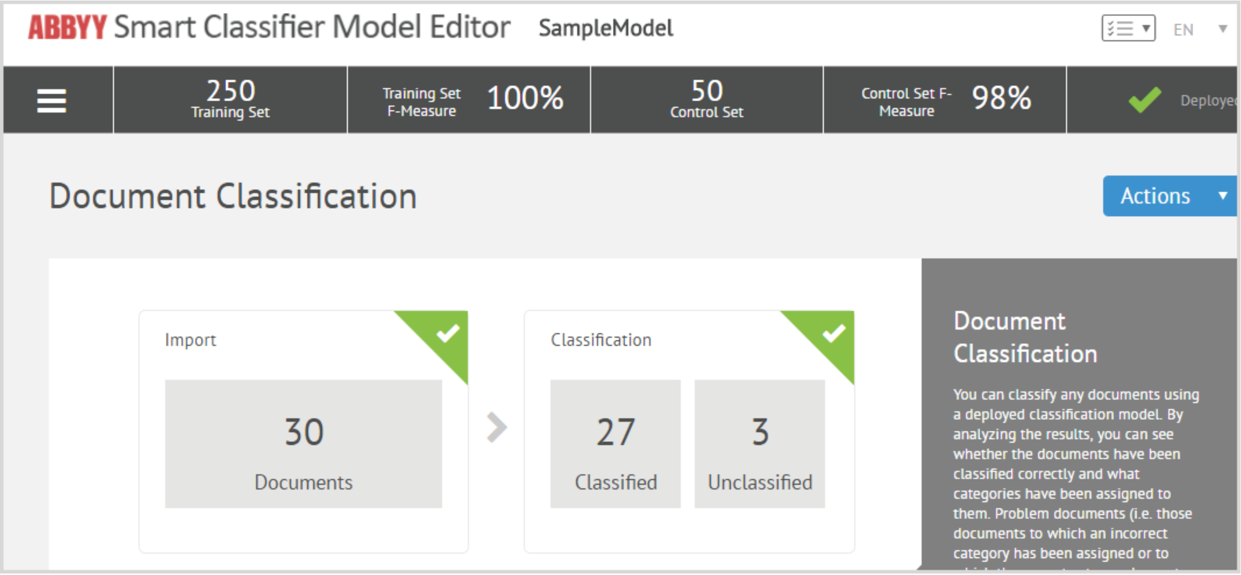
Documents can be classified either using the ABBYY Compreno REST API (the most commonly used approach; see the "Classifying documents and getting the results" section of the **ABBYY Compreno Products 2.7 Integration Guide**) or the ABBYY Smart Classifier Model Editor.

Only a deployed model can be used to classify documents in the ABBYY Smart Classifier Model Editor. If you have a deployed model available to you, complete these steps to classify your documents:

1. On the **Deploy Model** tile, click the **Documents to classify** button.
2. On the **Document Classification** screen, click the **Import** tile, browse to the **Classification set** folder, and import the documents from this folder. You can find the **Classification set** folder in:

   %PUBLIC%\ABBYY\Compreno Products\2.7\Code Samples\SmartClassifierSampleApplication\SampleSets.
3. Click the **Classify** tile.

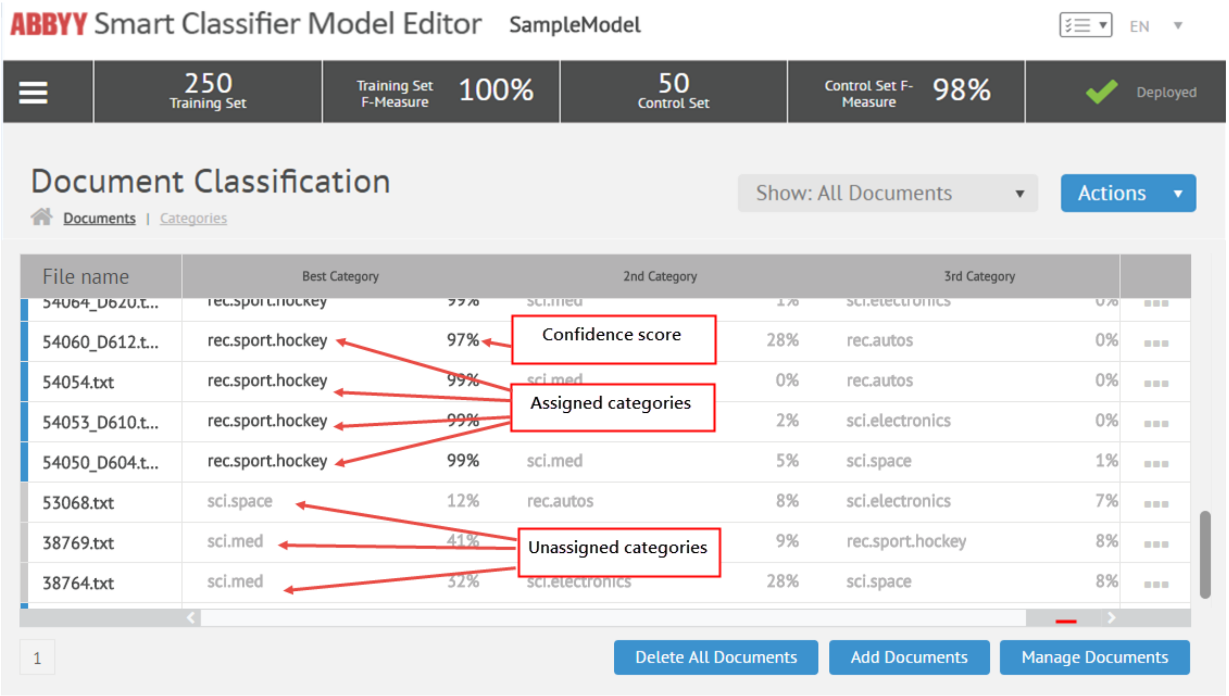
   After a while, the **Classification** tile will display the number of classified documents.



Document Classification

You can classify any documents using a deployed classification model. By analyzing the results, you can see whether the documents have been classified correctly and what categories have been assigned to them. Problem documents (i.e. those documents to which an incorrect category has been assigned or to

4. To see the categories assigned to each document, click the **Classified** button on the **Classification** tile.



ABBYY Smart Classifier Model Editor
SampleModel
EN
250 Training Set
Training Set F-Measure 100%
50 Control Set
Control Set F-Measure 98%
Deployed

Document Classification

File name	Best Category		2nd Category		3rd Category		
54064_D620.t...	rec.sport.hockey	99%	sci.med	1%	sci.electronics	0%	▪▪▪
54060_D612.t...	rec.sport.hockey	97%	[illegible]	28%	rec.autos	0%	▪▪▪
54054.txt	rec.sport.hockey	99%	sci.med	0%	rec.autos	0%	▪▪▪
54053_D610.t...	rec.sport.hockey	99%	[illegible]	2%	sci.electronics	0%	▪▪▪
54050_D604.t...	rec.sport.hockey	99%	sci.med	5%	sci.space	1%	▪▪▪
53068.txt	sci.space	12%	rec.autos	8%	sci.electronics	7%	▪▪▪
38769.txt	sci.med	41%	[illegible]	9%	rec.sport.hockey	8%	▪▪▪
38764.txt	sci.med	32%	sci.electronics	28%	sci.space	8%	▪▪▪

Usage Scenarios

This section describes several scenarios for using ABBYY Smart Classifier. We recommend that you start working with ABBYY Smart Classifier by selecting the scenario that best meets your needs.

- Handling requests from members of the public
- Analyzing technical support requests
- Selecting storage policies for documents
- Assigning attributes to documents

Handling requests from members of the public

In this scenario, ABBYY Smart Classifier is used to automate the classification of requests, complaints, petitions, appeals, etc. submitted by members of the public. Classified documents can then be automatically routed to the departments responsible for handling the issues in question. It is highly undesirable to have any of the incoming documents routed to the wrong departments, because operators in those departments will have to spend their time finding the correct addressee instead of dealing with their own documents.

Before you start implementing this scenario, compile a list of categories that can be assigned to such documents. Each category should correspond to one of the departments responsible for handling a particular issue.

To implement this scenario, complete the following steps:

1. Create, train, and deploy a classification model using the ABBYY Smart Classifier Model Editor (included in the distribution kit).

   Note: For this scenario, we recommend prioritizing precision over recall in the project settings. This will reduce the number of misclassified documents ending up in the wrong department. However, this will also increase the number of documents that will not be classified automatically and so will require manual classification.

2. Using the ABBYY Compreno REST API, get a list of categories.
3. By default, the list of categories contains all categories with high category confidence scores.

   If the list of categories you get contains a single reliably assigned category, use this category. If the list contains several categories, you can either:

- select the category with the highest confidence score (note, however, that this may increase the number of documents to which this category will be wrongly assigned) or
- examine the list and assign the correct categories manually.

4. Any documents to which no category has been assigned should be placed into a separate folder for subsequent manual classification.

Analyzing technical support requests

The main objective of this scenario is to find solutions to problems reported by users. The system analyzes the contents of a user's support question and returns a list of knowledge base articles that are most likely to contain the solution. The list of articles thus obtained can be displayed either to the user or to a technical support specialist.

Before you start implementing this scenario:

1. Compile a list of categories that can be assigned to such documents. Each category should correspond to a knowledge base article.
2. Prepare a set of documents pre-sorted into the right categories for training ABBYY Smart Classifier. Use resolved support tickets where solutions were found in specific knowledge base articles.

To implement this scenario, complete the following steps:

1. Create, train, and deploy a classification model using the ABBYY Smart Classifier Model Editor (included in the distribution kit).

   Note: For this scenario, we recommend prioritizing recall over precision in the project settings. This will reduce the number of unclassified tickets and return more knowledge base articles for each ticket. This will also increase the number of tickets classified incorrectly, but this is acceptable in this scenario.

2. Using the ABBYY Compreno REST API, get a list of categories.
3. After your documents are classified, use some of the categories with high category confidence scores.

Selecting storage policies for documents

In this scenario, ABBYY Smart Classifier is used to classify documents and incoming messages in order to determine their nature and importance so that a fitting storage policy can be selected for them. This scenario can be used to automatically remove no-longer-needed or outdated information.

Before you start implementing this scenario:

1. Compile a list of categories that can be assigned to such documents. Each category should correspond to a document storage policy. The categories below may serve as an example:
   - Important
     Documents in this category must never be deleted.
   - Relevant
     Documents in this category will not be deleted for a long period of time.
   - Miscellaneous
     Documents in this category will be regularly deleted.
2. Prepare a set of documents pre-sorted into the right categories for training ABBYY Smart Classifier.

   Note: Each category must include at least 10 documents.

To implement this scenario, complete the following steps:

1. Create, train, and deploy a classification model using the ABBYY Smart Classifier Model Editor (included in the distribution kit).
2. Using the ABBYY Compreno REST API, get a list of categories.
3. By default, the list of categories contains all categories with high confidence scores.

   If the list of categories you get contains a single reliably assigned category, use this category. If the list contains several categories, you can either:

   - select the category with the highest confidence score (note, however, that this may increase the number of documents to which this category will be wrongly assigned) or
   - examine the list and assign the correct categories manually.
4. Any documents to which no category has been assigned should be placed into a separate folder for subsequent manual classification.

Assigning attributes to documents

In this scenario, attributes are assigned to each document in a collection to facilitate document search. For example, news items may have such attributes as BUSINESS, POLITICS, SPORT, etc. A document may have more than one attribute.

Before you start implementing this scenario, compile a list of categories that can be assigned to such documents. Each category should correspond to an attribute.

To implement this scenario, complete the following steps:

1. Create, train, and deploy a classification model using the ABBYY Smart Classifier Model Editor (included in the distribution kit).

   Note: For this scenario, we recommend prioritizing recall over precision in the project settings. This will reduce the number of unclassified documents. This will also increase the number of documents classified incorrectly, but this is acceptable in this scenario.

2. Using the ABBYY Compreno REST API, get a list of categories.

3. When your documents are classified, use several categories with the highest confidence scores.

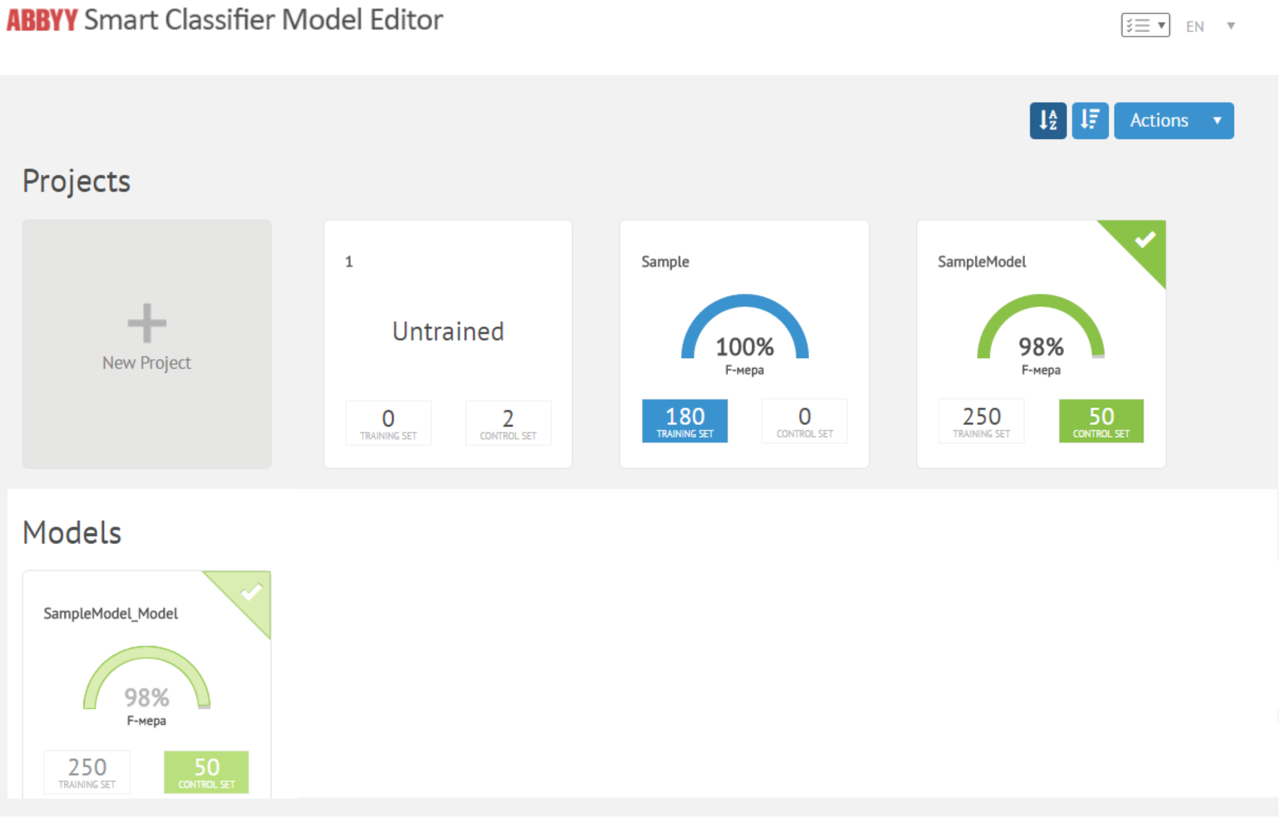
4. Any documents to which no category has been assigned should be placed into a separate folder for subsequent manual classification.

Creating a Classification Model

To start the ABBYY Smart Classifier Model Editor, click **Start > ABBYY Compreno Products > ABBYY Smart Classifier Model Editor**. You will see the main page of the editor.

In the top right corner, you can select Russian or English as the GUI language of the editor.

Classification models can be stored either within projects or separately. Consequently, in the editor, you will find your models either in the **Projects** section or in the **Models** section.

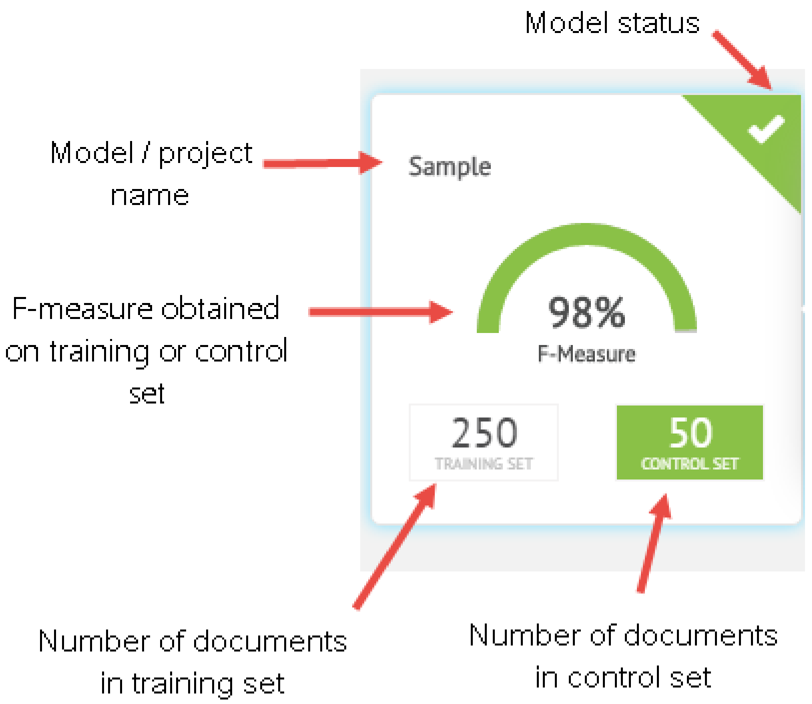


Before you can create a model, you must first create a project. A project contains a training set of documents, a control set of documents, and a trained classification model. The training, assessment, fine-tuning, and deployment of a model are all performed within a project.

The **Models** section displays the models that were extracted from projects by importing the respective .ascproj files. Imported models do not contain any documents for training or assessment, and so cannot be modified. For the same reason, imported models are much smaller than their projects.

Each project that you create and each model that you import are shown as tiles on the main page of the editor.

A project tile will look like this:


Model status
Model / project name
Sample
F-measure obtained on training or control set
98%
F-Measure
250
TRAINING SET
50
CONTROL SET
Number of documents in training set
Number of documents in control set

The colors in the tiles have the following meanings:

- Blue means that the model has been trained but has not yet been deployed.
- Green means that the model has been deployed.
- Yellow means that the model has been modified after deployment. As a result, the model exists in two versions — the original deployed version and the current modified version. When you point your mouse cursor to such a model, an information panel appears. Clicking this panel will flip the tile, allowing you to switch between the current and the deployed version.

Click the Actions ▾ button on the main page of the editor to see a menu of commands for working with projects and models. You can:

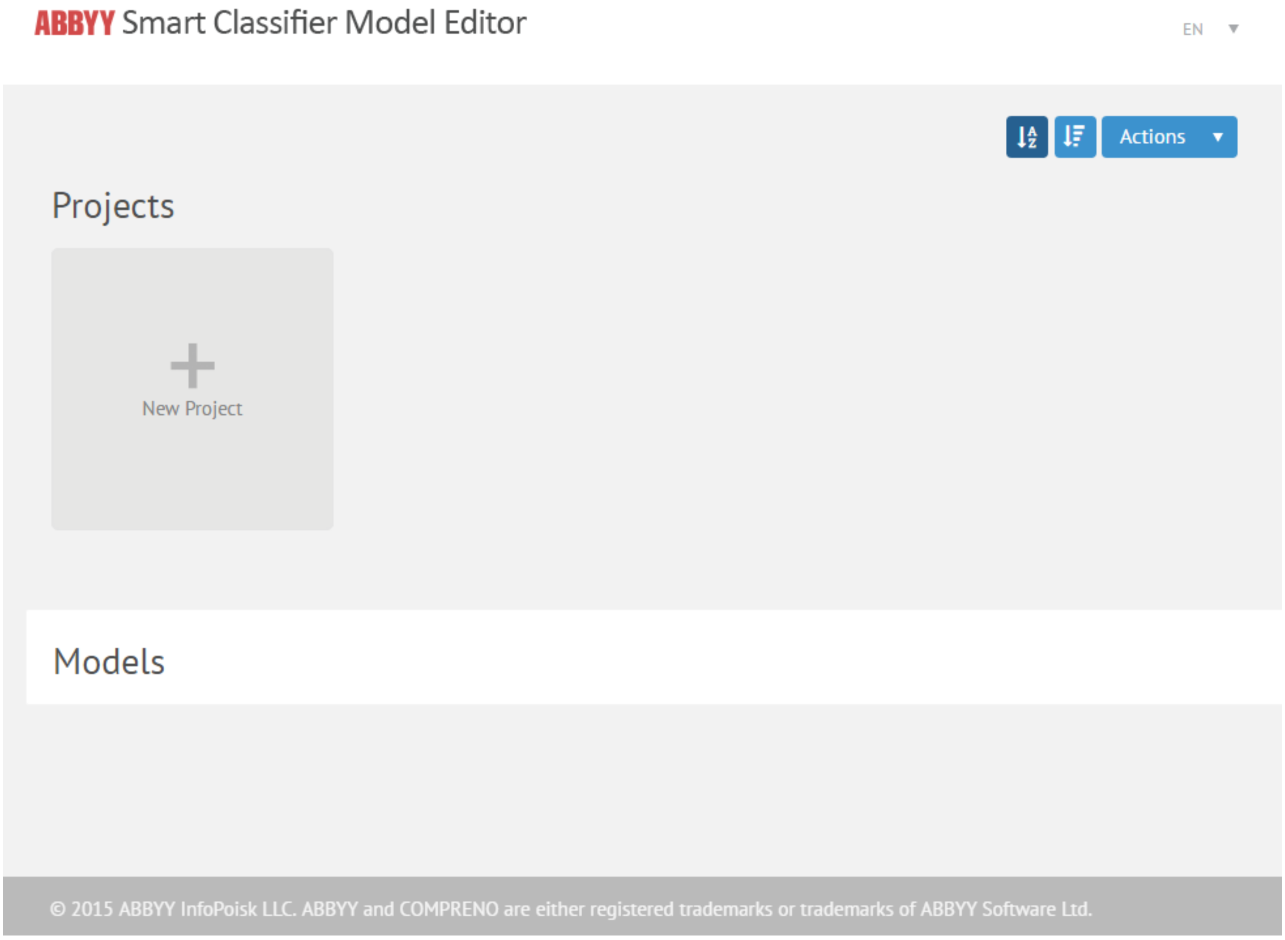
- Import a model or project.
- Export a model or project.
- Export a training or control set of documents.
- Export a category folder structure.
- Rename a model or project.
- Copy a model or project.
- Delete a model or project.

To obtain a working classification model, you need to complete the steps that make up the life cycle of a model.

Creating a project

To create a project, complete these steps:

1. On the main page of the ABBYY Smart Classifier Model Editor, click the **New Project** tile.


ABBYY Smart Classifier Model Editor
EN ▼
Actions ▼
Projects
New Project
Models
© 2015 ABBYY InfoPoisk LLC. ABBYY and COMPRENO are either registered trademarks or trademarks of ABBYY Software Ltd.

2. On the screen that opens, specify the following:

   - Provide a name for your project (this name will be used as the Classification.ModelName parameter when classifying documents by means of the REST API).

   - Specify the language of the documents that are to be classified. You will not be able to change the selected language after the project is created.

     Important! See the Appendix for a list of supported classification languages.

     If your document collection contains documents in multiple languages, create a separate project for each language (see Less Common Usage Scenarios for details).

   - Select a category assignment method. You will be able to change the category assignment method after training and/or assessing the model.

     The following options are available:

- ○ **Single candidate category**

  Use this option in cases where assigned categories will not be checked manually and assigning the wrong category is highly undesirable.

- ○ **Top candidate category**

  Use this option in cases where each document can belong to only one category and classification errors can be tolerated.

- ○ **All candidate categories**

  Use this option if a document may belong to multiple categories or if the correct category will be selected manually by verifying candidate categories.

3. Click the **Next** button.

4. In the next screen, use the **Inclusiveness** slider to specify a desired recall-to-precision ratio (if you know your classification priorities) or use the default **BALANCED** setting.

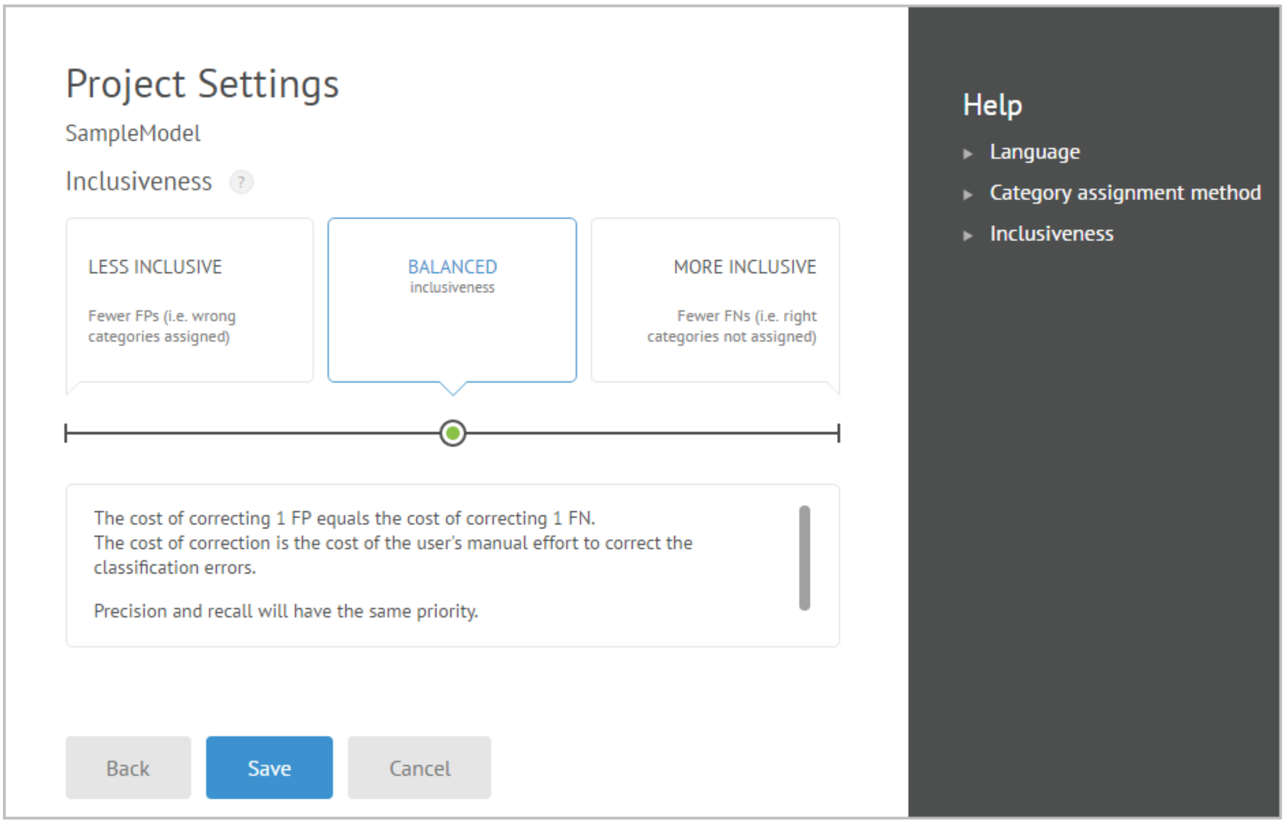
   You will be able to modify this setting after you create your project.

   If you give priority to precision over recall, the results will contain fewer false positives, i.e. cases when the wrong category has been assigned to a document. This may be a reasonable choice when, for example, you classify requests from members of the public and it is highly desirable to have each request routed to the right department.

   If you give priority to recall over precision, the results will contain fewer false negatives, i.e. cases when the right category has not been assigned to a document. This may be a reasonable choice when, for example, you classify incoming technical support requests into different categories. There won't be much harm if some requests have redundant categories assigned to them, but it is highly undesirable not to have the right category assigned to a user's request.

   ❗**Note:** *Please note that the inclusiveness setting only affects the choice of the best model at the training stage, when the program selects the model that keeps the cost of error correction to a minimum (i.e. any other model would require more human effort to correct classification errors). The ratio of FPs to FNs in a training/assessment report will not be the same as that selected by means of the inclusiveness slider.*

   For more tips on selecting the right **Inclusiveness** slider setting for your project, see Improving Your Classification Model.



Project Settings

SampleModel

Inclusiveness

?
LESS INCLUSIVE
Fewer FPs (i.e. wrong categories assigned)
BALANCED
inclusiveness
MORE INCLUSIVE
Fewer FNs (i.e. right categories not assigned)

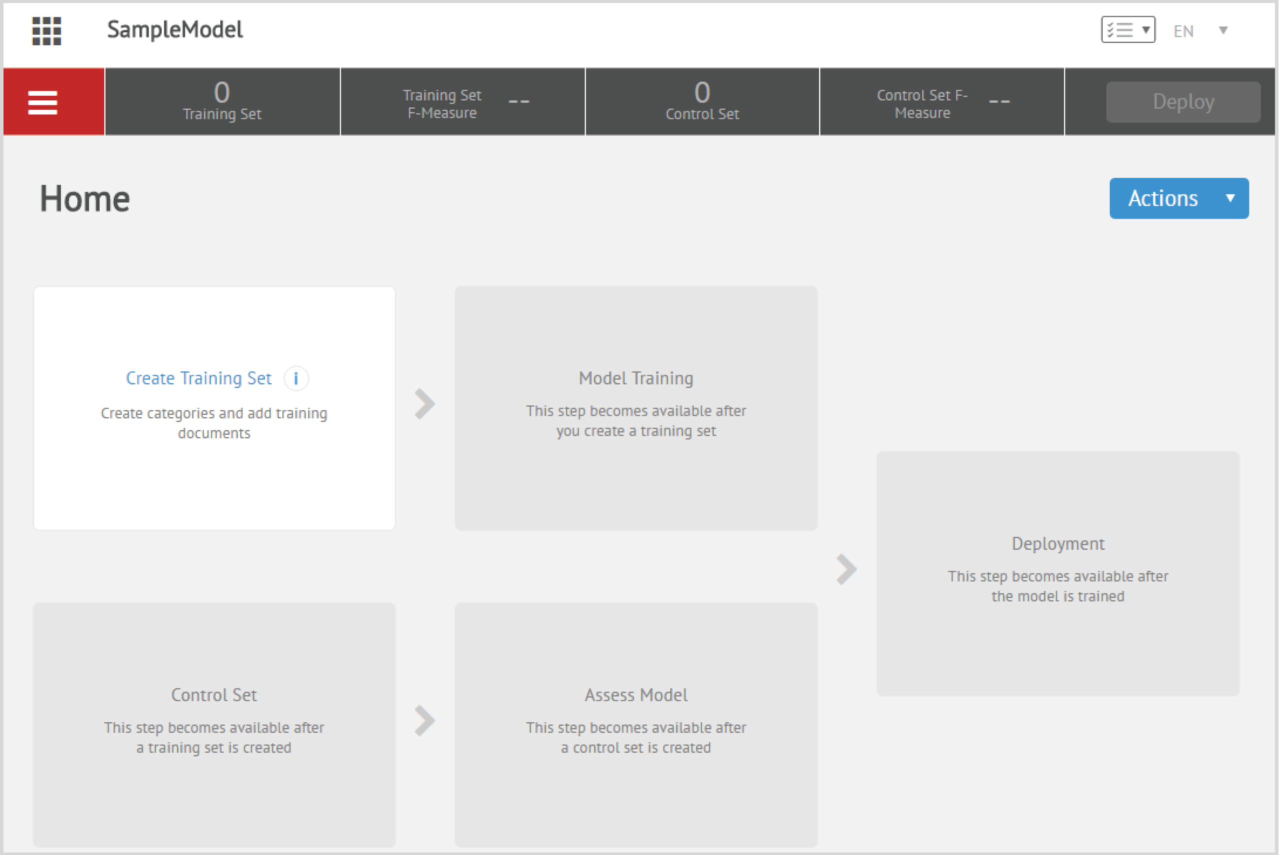
The cost of correcting 1 FP equals the cost of correcting 1 FN.
The cost of correction is the cost of the user's manual effort to correct the classification errors.

Precision and recall will have the same priority.

Back
Save
Cancel
Help
Language
Category assignment method
Inclusiveness

5. Click the **Save** button.

The home page of your newly created project will open, where you can see the steps to be performed in order to obtain a working classification model. Click the ⓘ icon in each tile to see a detailed description of the respective step.



Home

Create Training Set

Create categories and add training documents

Model Training

This step becomes available after you create a training set

Deployment

This step becomes available after the model is trained

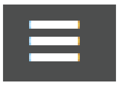
Control Set

This step becomes available after a training set is created

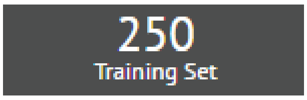
Assess Model

This step becomes available after a control set is created

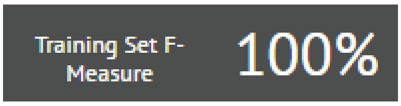
At the top of the project home page, you can see a bar containing the following tabs:



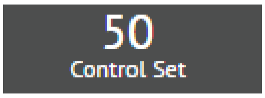
Clicking this icon will take you to the home page, and placing your mouse cursor over it will open a menu of commands.



This tab shows the number of documents in the training set. Click this tab to see the categories in the training set.



This tab displays the F-measure obtained for the model after it has been trained on the training set.
Click this tab to see the training statistics for the model.



This tab shows the number of documents in the control set. Click this tab to see the categories in the control set.

This tab displays the F-measure obtained for the model after it has been assessed using the control set.
Click this tab to see the assessment statistics for the model.

This button becomes available after you have trained or assessed your model.
Click this button to deploy your model. Once the model is deployed, the tab will change to .

Creating a training set

To train your classification model, you need to create a training set of documents.

A training set should contain documents assigned to their appropriate categories.

For example, your classification project may contain such categories as CONTRACT, POLICY, and CERTIFICATE, and the CONTRACT category may be assigned to commercial contracts, license agreements, etc.

The quality of a classification model depends on the quality of the training set on which it was trained. When creating a training set, keep the following in mind:

- You need a representative sample of documents, i.e. the portion of documents assigned to each category must be roughly the same as in the document collection that you want to classify.

  For example, if about 70% of documents in your newswire collection will be about sport, then 70% of documents in your training set should fall into the SPORT category.

- Your training set must contain at least 2 categories, each containing at least 10 documents.
- It is recommended to have at least 100 documents in each category to ensure that the program selects the optimal classification algorithm.
- A category name may not be more than 255 characters long.
- A category name may not contain any of the following characters: **#**, **@**, **%**, **&**, ****, **/**, **:**, *****, **?**, **"**, **<**, **>**, **|**.
- A category name may not end with a dot.
- For unformatted text files (*.txt), Unicode or UTF-8 with BOM are recommended.
- It is not recommended to have more than 2,000 categories in a training set.

There are two ways to create a training set in the ABBYY Smart Classifier Model Editor:

- importing a ZIP archive
- adding folders and documents manually

Importing a ZIP archive

Important! The program will not accept ZIP archives larger than 2 GB. If the zipped version of your training set is greater than 2 GB, split it into smaller fragments and import them one by one.

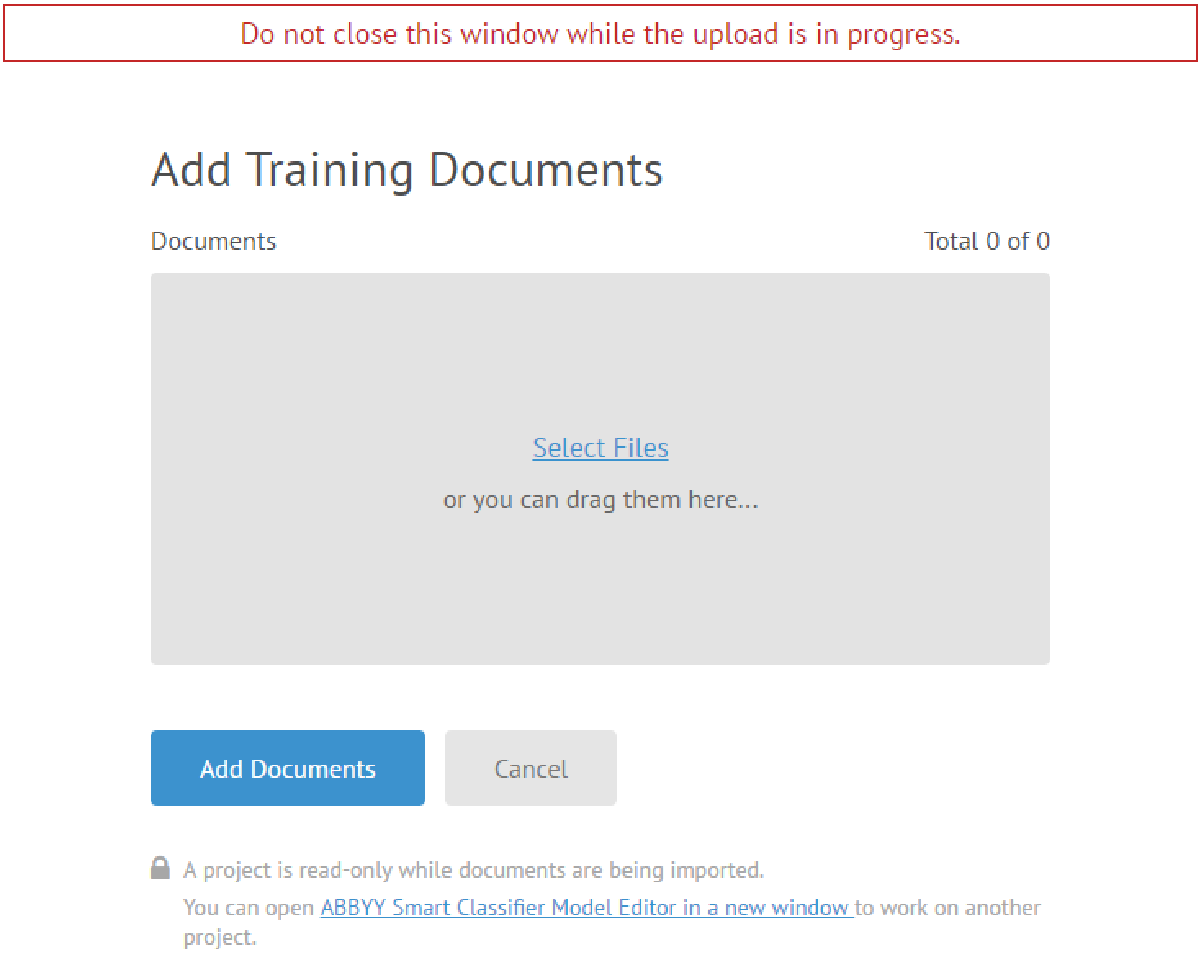
After you import the training set, training will start automatically. To disable automatic training, clear the option **Start training automatically once the ZIP archive is imported.**

Adding documents manually

Alternatively, you can create a folder structure corresponding to your categories and subcategories directly in the ABBYY Smart Classifier Model Editor and then put each document in its appropriate folder.

To add documents, complete the steps below:

1. Click the **Training Set** tile. A page will open displaying the list of categories in the training set.
2. To add documents to a category, click it and then click the Add Documents button and select the files you wish to add. In the dialog box that opens, add the documents.



Do not close this window while the upload is in progress.

Add Training Documents

Documents

Total 0 of 0

Select Files

or you can drag them here...

A project is read-only while documents are being imported.

You can open ABBYY Smart Classifier Model Editor in a new window to work on another project.

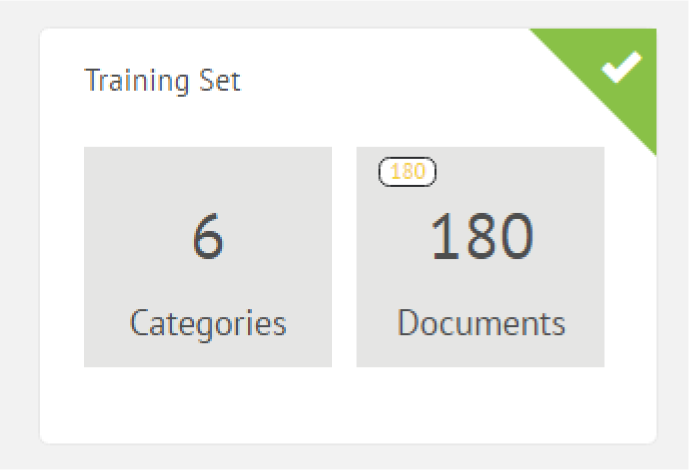
.

The selected documents will appear in the **Add Training Documents** window.

3. Click the **Import** button and keep the window open while the documents are being imported.

The selected documents will be imported into the training set of your project. The total number of documents will be displayed on the **Training Set** tab at the top of the screen.

On the home page of the project, you will see a **Training Set** tile with a green dog ear (if import was successful). The number of errors and warnings that occurred when the project's documents were imported will be displayed in the upper-left corner of the **Documents** box. Click the number to view the categories and documents with errors and warnings and decide whether the impact of these errors and warnings on classification is significant enough to require fixing (see *Improvements at the category level* and *Improvements at the document level* for details). Once you determine that there are no significant errors and warnings, you can train your classification model.


Training Set

6

Categories

180

180

Documents

Training Your Classification Model

When you train your model, the program determines:

- What features are typical of each category.
- What algorithm should be used to distinguish documents of one category from documents of the other categories contained in the training set.

Once the training is complete:

- A classification model is created that provides the best quality of classification that can be achieved for the training set of documents.
- The documents in the training set are classified using the trained classification model.

You cannot objectively assess the quality of your model based on the classification results you obtain on a training set of documents. The quality of a model should be assessed using a control set of documents.

When you train your model, the project switches to read-only mode, i.e. it cannot be modified until the training is finished.

Training times depend on the number of categories and documents in the training set.

By default, training starts automatically after you import the training set. If you disabled the automatic start earlier, click the **Training** tile after importing the training set.

Classification statistics

Once the training is complete, the following statistics are provided for each document and for each true category:

- If a true category is assigned to a document, this fact is logged as a true positive (TP). A true positive will occur, for example, if the SPAM category is assigned to a spam e-mail message.
- If a category not assigned to a document is not a true category, this fact is logged as a true negative (TN). A true negative will occur, for example, if the SPAM category is not assigned to a legitimate e-mail message.
- If a category assigned to a document is not a true category, this fact is logged as a false positive (FP). A false positive will occur, for example, if the SPAM category is assigned to a legitimate e-mail message.
- If a true category is not assigned to a document, this fact is logged as a false negative (FN). A false negative will occur, for example, if the SPAM category is not assigned to a spam e-mail message.

On the statistics page, you can see how many times a particular category:

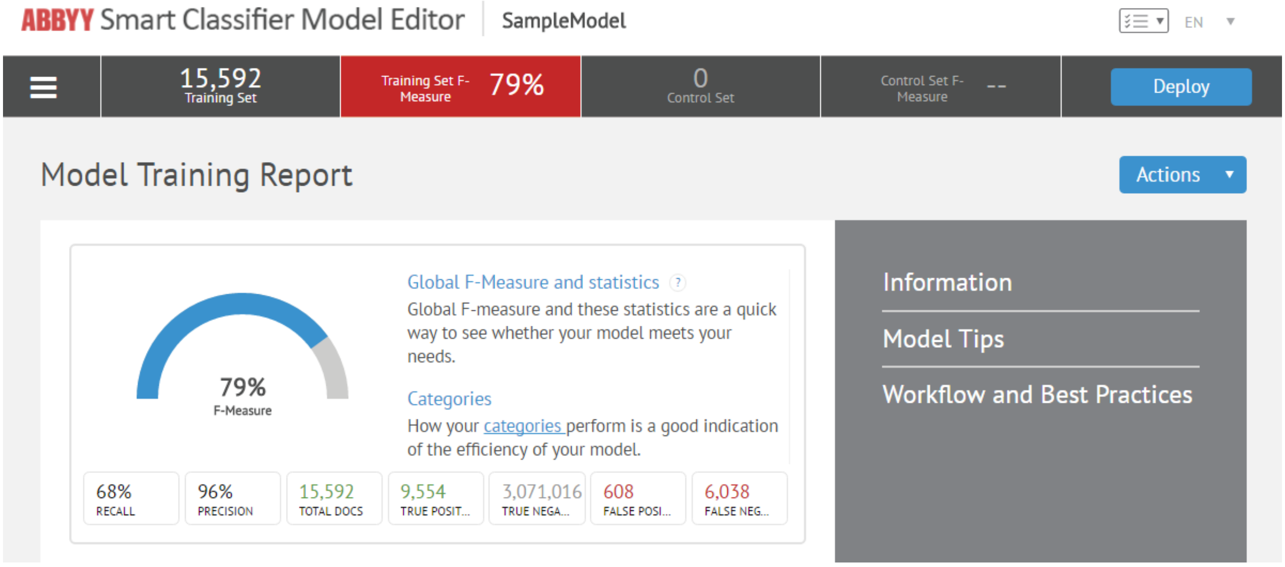
- has been assigned correctly (TPs)
- rightly, has not been assigned (TNs)
- has been assigned incorrectly (FPs)
- wrongly, has not been assigned (FNs)

You can also see the total numbers of TPs, TNs, FPs, and FNs obtained with the model.

The sum total of all of the TPs, TNs, FPs, and FNs may not equal the total number of documents. To see why, consider this example. Suppose there are 10 categories in your classification project and each document in your control set has one true category. If you choose **All Candidate Categories** as the category assignment method, then several different categories may be assigned to each document. Now suppose that during classification, three categories have been assigned to one of the documents and only one of them is a true category. This means that for this document alone there have occurred 1 TP, 2 FPs, 7 TNs, and 0 FNs, so summing the TPs, TNs, FPs, and FNs for all of the documents will give you a number that will be greater than the number of documents.

Training report

Once the training is finished, you can see a training report to assess the performance of your model. To view the report, click the **Training Set F-Measure** tab at the top of the screen. The report contains the following metrics: F-measure, TPs, TNs, FPs, and FNs obtained for the training set.



Model Training Report

Global F-Measure and statistics

Global F-measure and these statistics are a quick way to see whether your model meets your needs.

Categories

How your categories perform is a good indication of the efficiency of your model.

68%	96%	15,592	9,554	3,071,016	608	6,038
RECALL	PRECISION	TOTAL DOCS	TRUE POSIT...	TRUE NEGA...	FALSE POSI...	FALSE NEG...

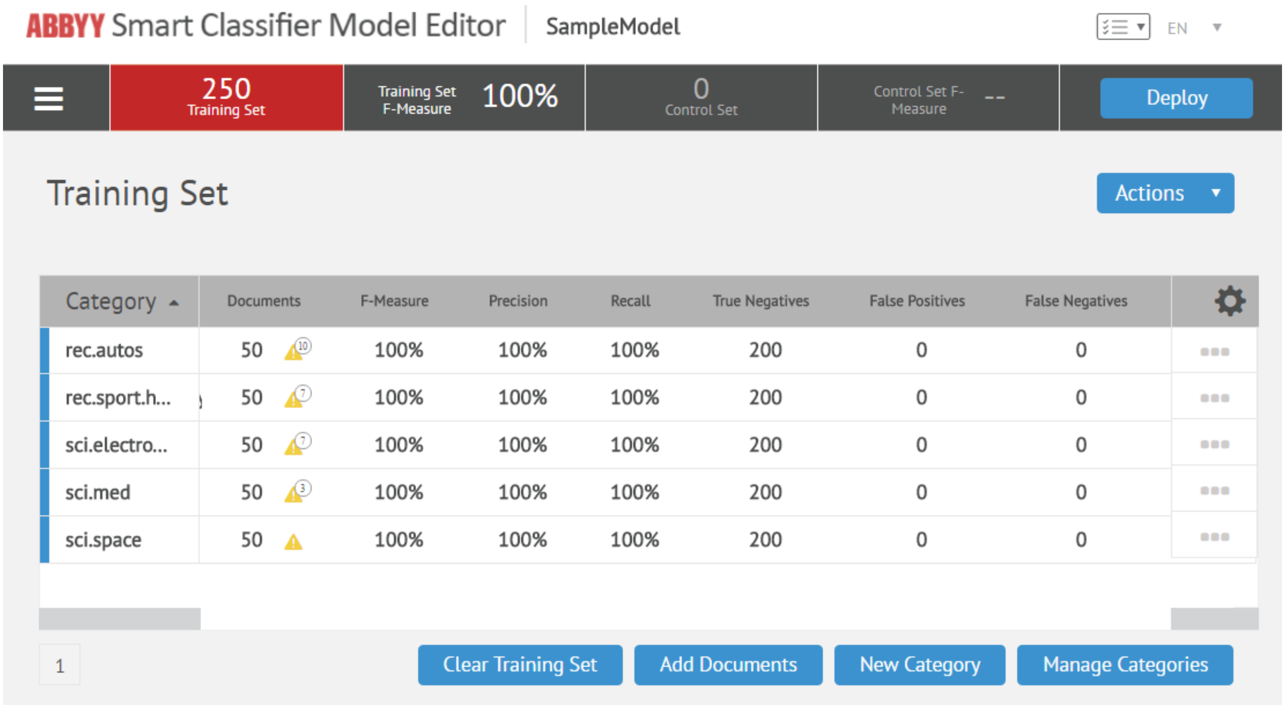
Obviously, greater precision and recall values mean better classification results. The weighted average of precision and recall is termed F-measure. For some tips on improving the F-measure, see Improving Your Classification Model.

The statistics are generated from the metrics obtained for each category.

Statistics for categories and documents

You can review the classification quality achieved for individual categories. Click the **Training Set** tab at the top of the screen to open the page with the document categories. This page shows the categories available in your training set and the number of documents in each category. After you train or assess your model, additional statistics become available.

To see the number of TPs, TNs, FPs, and FNs for each category, click the ⚙ icon and select the items to display.



Training Set

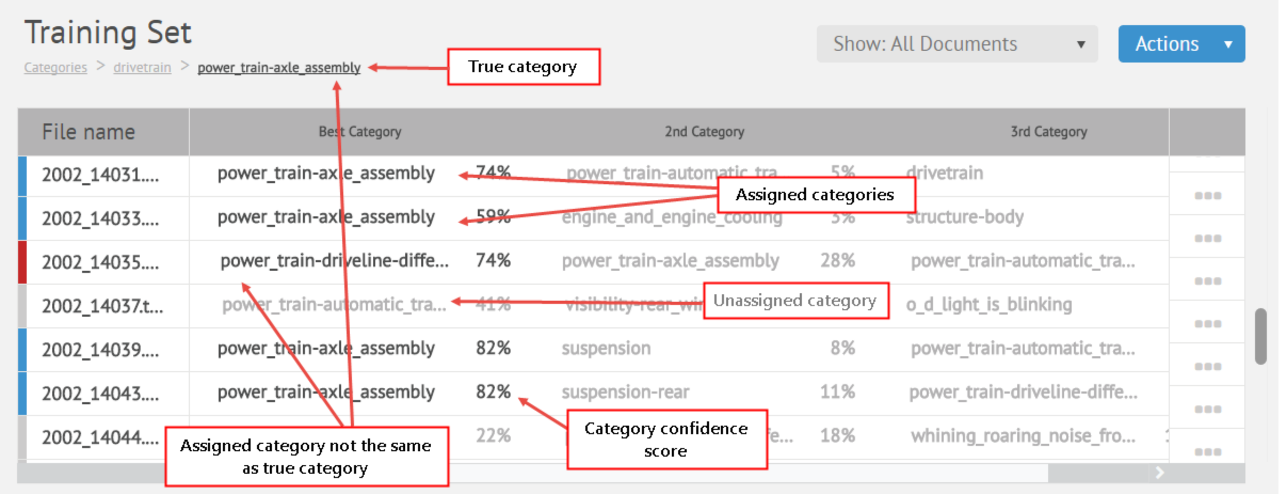
Category	Documents	F-Measure	Precision	Recall	True Negatives	False Positives	False Negatives
rec.autos	50 ⚠10	100%	100%	100%	200	0	0
rec.sport.h...	50 ⚠7	100%	100%	100%	200	0	0
sci.electro...	50 ⚠7	100%	100%	100%	200	0	0
sci.med	50 ⚠3	100%	100%	100%	200	0	0
sci.space	50 ⚠	100%	100%	100%	200	0	0

The icons next to the numbers in the **Documents** column have the following meanings:

⚠10 indicates that the category contains documents with warnings (the number is the number of such documents).

⚠2 indicates that the category contains documents with errors (the number is the number of such documents).

Open a category to see a list of documents with classification results obtained for each document.



Training Set

Categories > drivetrain > power_train-axle_assembly

File name	Best Category		2nd Category		3rd Category
2002_14031....	power_train-axle_assembly	74%	power_train-automatic_tra...	5%	drivetrain
2002_14033....	power_train-axle_assembly	59%	engine_and_engine_cooling	5%	structure-body
2002_14035....	power_train-driveline-diffe...	74%	power_train-axle_assembly	28%	power_train-automatic_tra...
2002_14037.t...	power_train-automatic_tra...	41%	visibility-rear_wi...		o_d_light_is_blinking
2002_14039....	power_train-axle_assembly	82%	suspension	8%	power_train-automatic_tra...
2002_14043....	power_train-axle_assembly	82%	suspension-rear	11%	power_train-driveline-diffe...
2002_14044....		22%	...e...	18%	whining_roaring_noise_fro...

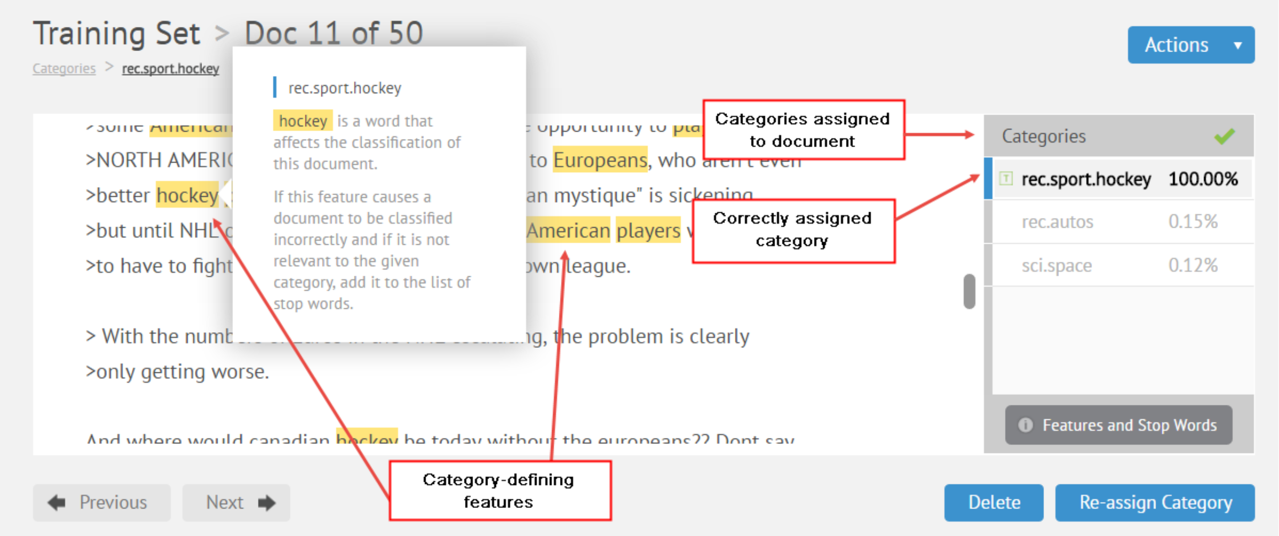
For each document, three candidate categories are displayed that have the highest category confidence scores. The font colors have the following meanings:

- Black indicates an assigned category.
- Gray indicates an unassigned category.

After the documents are classified, a vertical color bar will appear to the left of the name of each document. The bar colors have the following meanings:

- Blue indicates a correctly classified document.
- Gray indicates an unclassified document.
- Red indicates an misclassified document.

Clicking a document in the list will open the text of the document. The features that made the program place the document in this particular category will be highlighted in orange.



Training Set > Doc 11 of 50

Categories > rec.sport.hockey

rec.sport.hockey

hockey is a word that affects the classification of this document.

If this feature causes a document to be classified incorrectly and if it is not relevant to the given category, add it to the list of stop words.

>NORTH AMERIC … to Europeans, who aren't even

>better hockey … an mystique" is sickening

>but until NHL … American players

>to have to fight … own league.

> With the numb … ng, the problem is clearly

>only getting worse.

And where would canadian hockey be today without the europeans?? Dont say

Categories	✓
rec.sport.hockey	100.00%
rec.autos	0.15%
sci.space	0.12%

Actions ▾
Categories assigned to document
Correctly assigned category
Category-defining features
Features and Stop Words
Previous
Next
Delete
Re-assign Category

Classification statistics obtained on a training set should not be taken to mean that the same precision and recall values will be achieved when classifying documents in the future. To get a more objective assessment of your model, you should assess it using a control set of documents.

Once the training is complete, the **Train Model** tile on the home page of the project will have a green dog ear. This means that the training has been successful and you can now deploy your model.

Important! Always assess your model on documents different from those that were used for training.

Creating a control set

To be sure of the good quality of your classification model, you need to assess it on documents other than those included in the training set. In the ABBYY Smart Classifier Model Editor, a control set of documents is used for this purpose.

A control set contains documents to which their appropriate categories have been assigned by the user.

For correct evaluation, your control set must meet the following requirements:

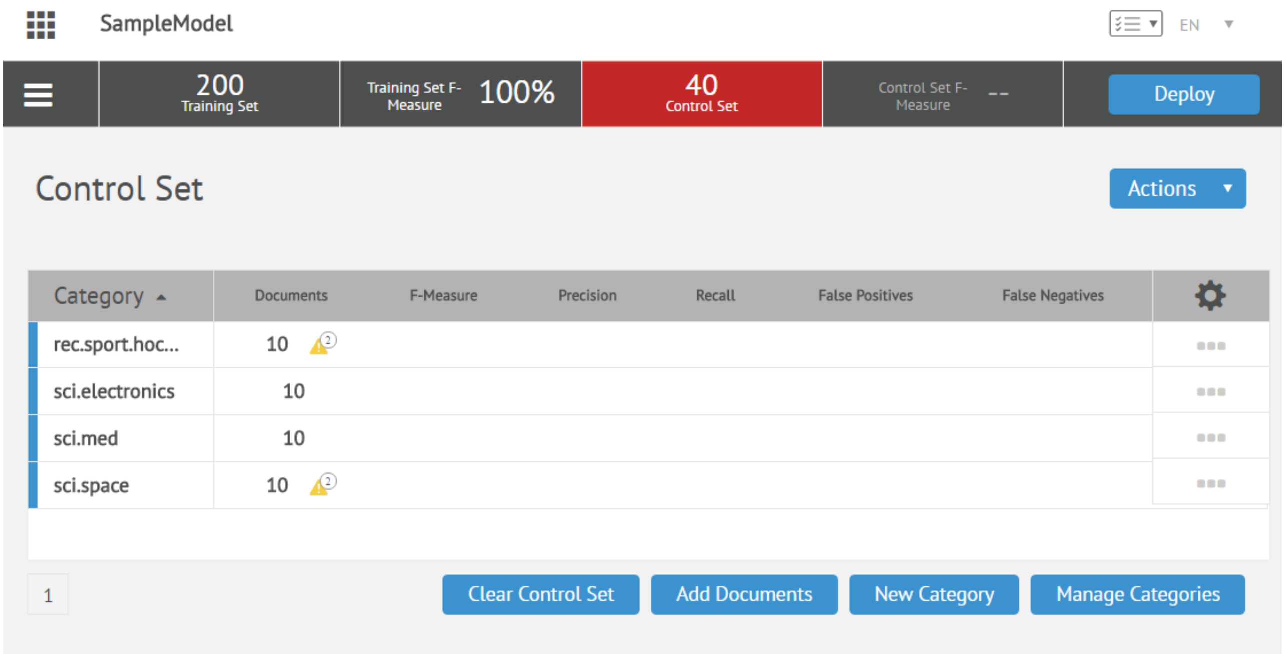
- You need a representative sample of documents, i.e. the portion of documents assigned to each category must be roughly the same as in the document collection that you want to classify.
- Minimal requirements:
  - The folder structure of the control set must be the same as the folder structure of the training set.
  - There must be at least one document in each category.
- Recommendations:
  - The control set should not contain documents from the training set.
  - It is recommended to have at least 100 documents in each category if you need to assess the classification power for each category.
  - It is recommended to have at least 100 documents in the control set if you need to assess the classification power of the entire model.

Creating a control set of documents

To create and import a control set, click the **Control set** tile and follow the instructions that appear on the screen.

Important! After you import the control set, assessment will start automatically. To disable automatic assessment, clear the option **Start the assessment automatically once import is complete**.

Once the documents are imported, the control set page will display the available categories and the number of imported documents.



Category ▲	Documents	F-Measure	Precision	Recall	False Positives	False Negatives
rec.sport.hoc...	10 ⚠2					
sci.electronics	10					
sci.med	10					
sci.space	10 ⚠2					

On the home page of the project, you will see a control set tile with a green dog ear (if import was successful) or with a yellow dog ear (if any warnings were issued). If, in your opinion, the warnings will have no effect on document classification, just ignore them. Now you can assess your classification model.

Assessing your model

Assessment is the process of classifying a control set of documents using a trained model. Assessment results allow you to evaluate the classification power of a trained model.

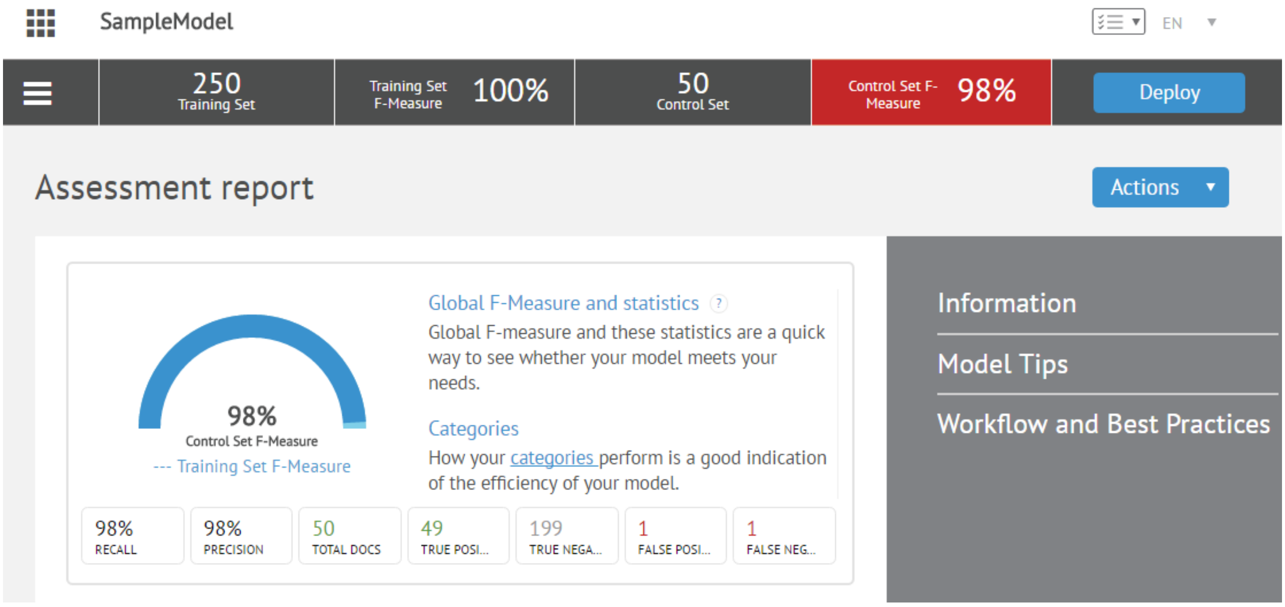
If the assessment did not start automatically after you imported the control set, click the **Assess Model using the Control Set** tile, and then click the **Assess** button.

Assessment times depend on the number of documents in the control set.

While model assessment is in progress, the project is read-only.

Once the assessment is complete, an assessment report will be generated. This report is an objective evaluation of the performance of your classification model, because the statistics in this report are obtained on documents other than those that were used for training. To see the report, click the **Control Set F-Measure** tab at the top of the screen. The F-measure calculated for the control set is an assessment of the classification power of your model, and you can safely assume that similar classification quality will be achieved on other documents in the future.

The following statistics are available in the report: F-measure, TPs, TNs, FPs, and FNs.



Global F-Measure and statistics

Global F-measure and these statistics are a quick way to see whether your model meets your needs.

Categories

How your categories perform is a good indication of the efficiency of your model.

RECALL	PRECISION	TOTAL DOCS	TRUE POSI...	TRUE NEGA...	FALSE POSI...	FALSE NEG...
98%	98%	50	49	199	1	1

You can see the classification quality achieved for individual categories. Click the **Control Set** tab at the top of the screen to open the page with the document categories.

If you need to improve your model, follow the tips in the Improving Your Classification Model section.

If you are satisfied with the trained model, deploy it.

Deploying your model

Complete the following steps to deploy your model:

1. Click the **Deploy** tile.
2. In the dialog box that opens, click the **Deploy Model** button.

The deployed model can be used for classifying documents either via the ABBYY Compreno REST API and on the **Document Classification** page.

For a detailed description of the REST API methods, please refer to the Integration Guide, which can be found by clicking **Start > ABBYY Compreno Products > Documentation > IntegrationGuide_English** on the computer where the ABBYY Compreno Products are installed.

Classifying documents

You can use the **Document Classification** page not only to classify documents, but also to find and correct classification errors. By classifying documents one by one, you can see why the program assigns a particular category to a particular document, analyze the classification errors that have occurred, and make improvements to your classification model.

To classify a document, complete the following steps:

1. On the home page of your project, click the **Documents to classify** button on the **Deploy Model** tile.
2. On the **Document Classification** page, click the **Import** tile and add some documents.

Classification will start automatically. The Classification tile will display the number of classified and unclassified documents.

To view the categories that have been assigned to the documents, click the **Classified** button on the **Classification** tile.

Analyze the classification results and adjust the settings or modify the training set if necessary. Then update and re-deploy your model and try classifying the documents again.

Improving Your Classification Model

Evaluating your model

Generally, classification models can be evaluated using the F-measure — the greater the F-measure, the better the quality of the model. Good F-measures are those which are close to F-measures obtained in manual classification.

When evaluating your model, consider the following:

- Your usage scenario

  In some usage scenarios, the F-measure may not be a good quality indicator if either precision or recall is more important.
  For example, when handling requests from members of the public, it is highly undesirable to have any requests routed to the wrong departments. In this scenario, precision is more important, so we need to reduce the number of requests classified incorrectly (i.e. reduce the number of FPs).
  On the other hand, when classifying technical support requests, it is acceptable to have redundant categories assigned to some requests, but absolutely unacceptable to have overlooked categories that should have been assigned. In this scenario, recall is more important, so we need to reduce the number of FNs.

  Use the F-measure in scenarios where both precision and recall are equally important.

- The size of your collection

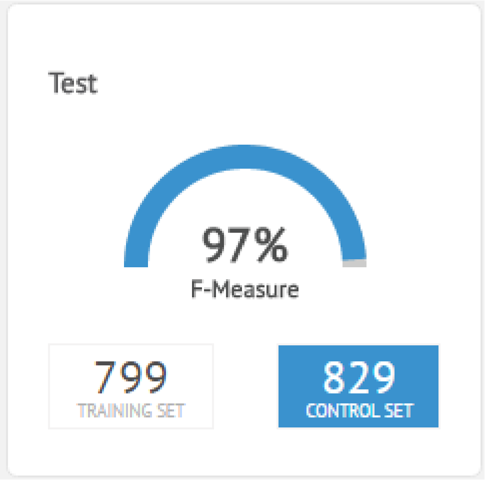
  When you manually classify, say, 60,000 documents into 150 categories, an F-measure of 60% may be considered a good result, because classification errors in this case are almost inevitable. Consequently, if for a similar number of documents classified with ABBYY Smart Classifier you get an F-measure greater than 60%, you can say that you have a created good classification model.

Reviewing the model metrics

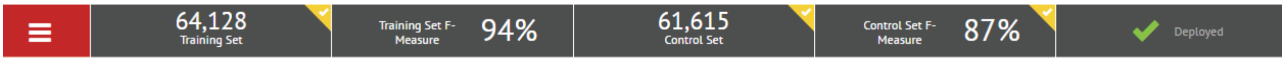
ABBYY Smart Classifier computes the F-measure, precision, and recall that can be obtained by using your classification model, first on a training set and then on a control set of documents. The values obtained on the control set are a more objective assessment of your model, because the documents in the control set have not been used in training.

You can see the F-measure:

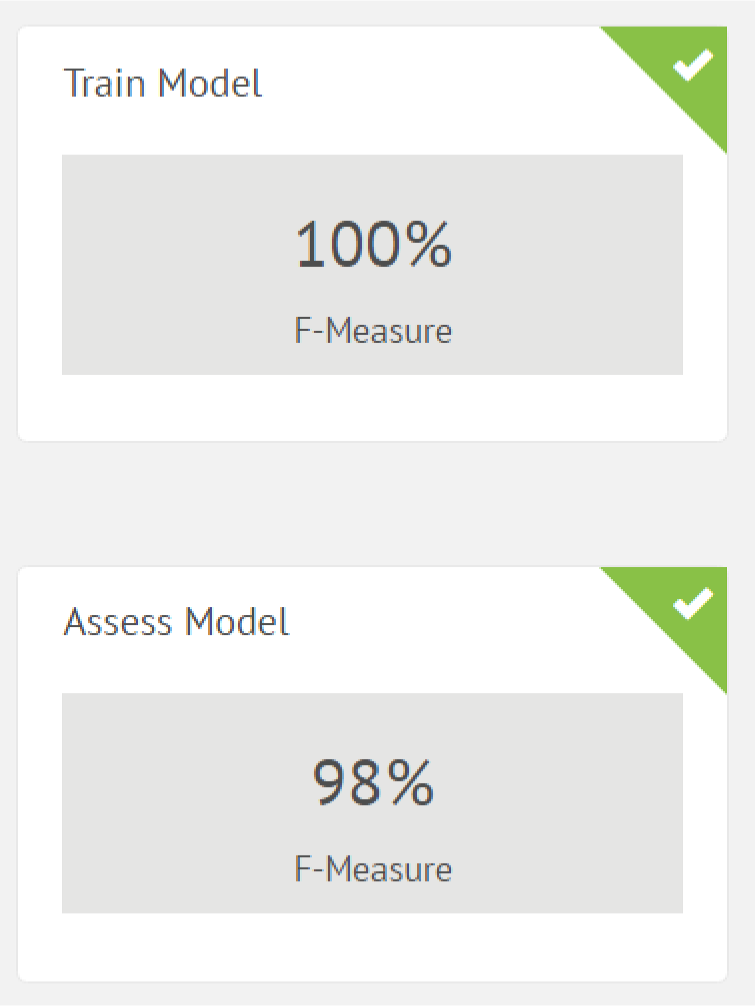
- on the tile of your project that appears on the main page



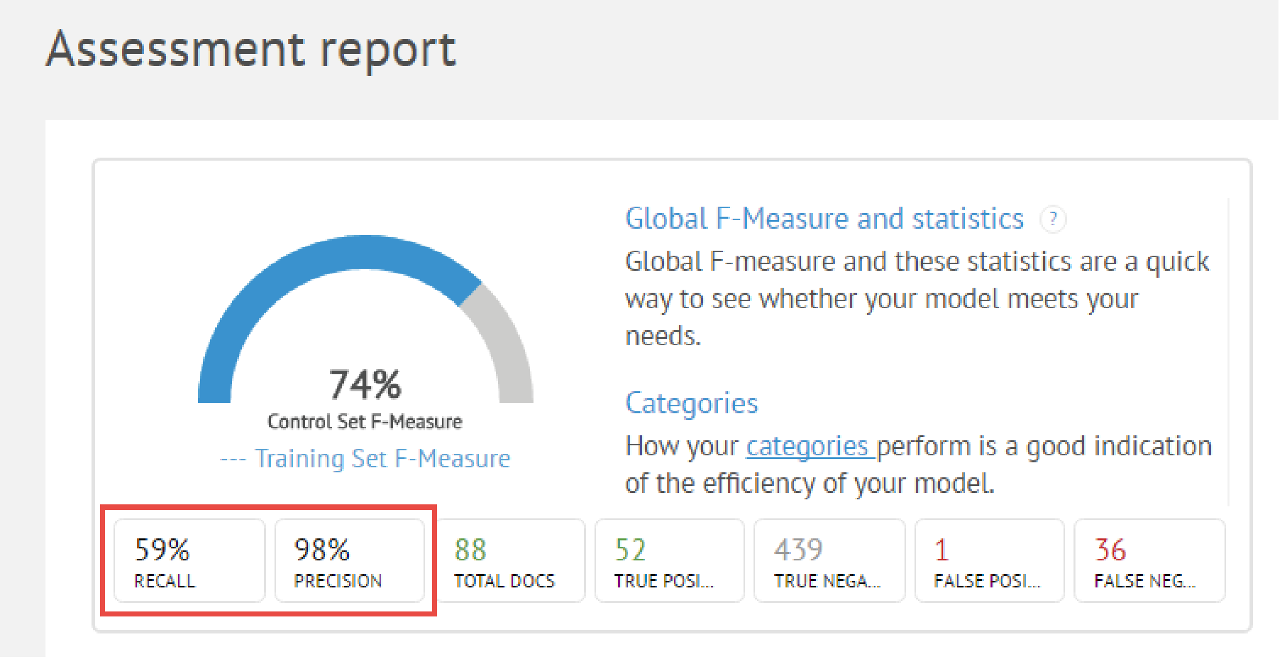
- on the bar at the top of the home page of the project



- on the **Train Model** and **Assess Model** tiles that appear on the home page of the project



The precision and recall values can be found in the training and assessment reports.



Assessment report

74%

Control Set F-Measure

--- Training Set F-Measure

Global F-Measure and statistics ?

Global F-measure and these statistics are a quick way to see whether your model meets your needs.

Categories

How your categories perform is a good indication of the efficiency of your model.

59%	98%	88	52	439	1	36
RECALL	PRECISION	TOTAL DOCS	TRUE POSI...	TRUE NEGA...	FALSE POSI...	FALSE NEG...

Problems and solutions

To see if your proposed adjustments will improve the quality of your model, create a copy of your project and make any adjustments that you think necessary to the models in this copy (e.g. for a deployed model named *Test* create copies named *Test_1*, *Test_2*, etc.). Then select the best model by comparing the results obtained with *Test_1*, *Test_2*, etc. against those obtained with the original project. Once you are satisfied with the results, deploy the best model. (If the original model is already integrated into your system as part of ABBYY Smart Classifier, you will need to change the name of the best adjusted model to that to that of the original model).

There are several things you can do to improve your model. You can make improvements at the following three levels of the project:

- Model
- Categories
- Documents

Important! After you make changes to the project settings or to the training set, be sure to update your model.

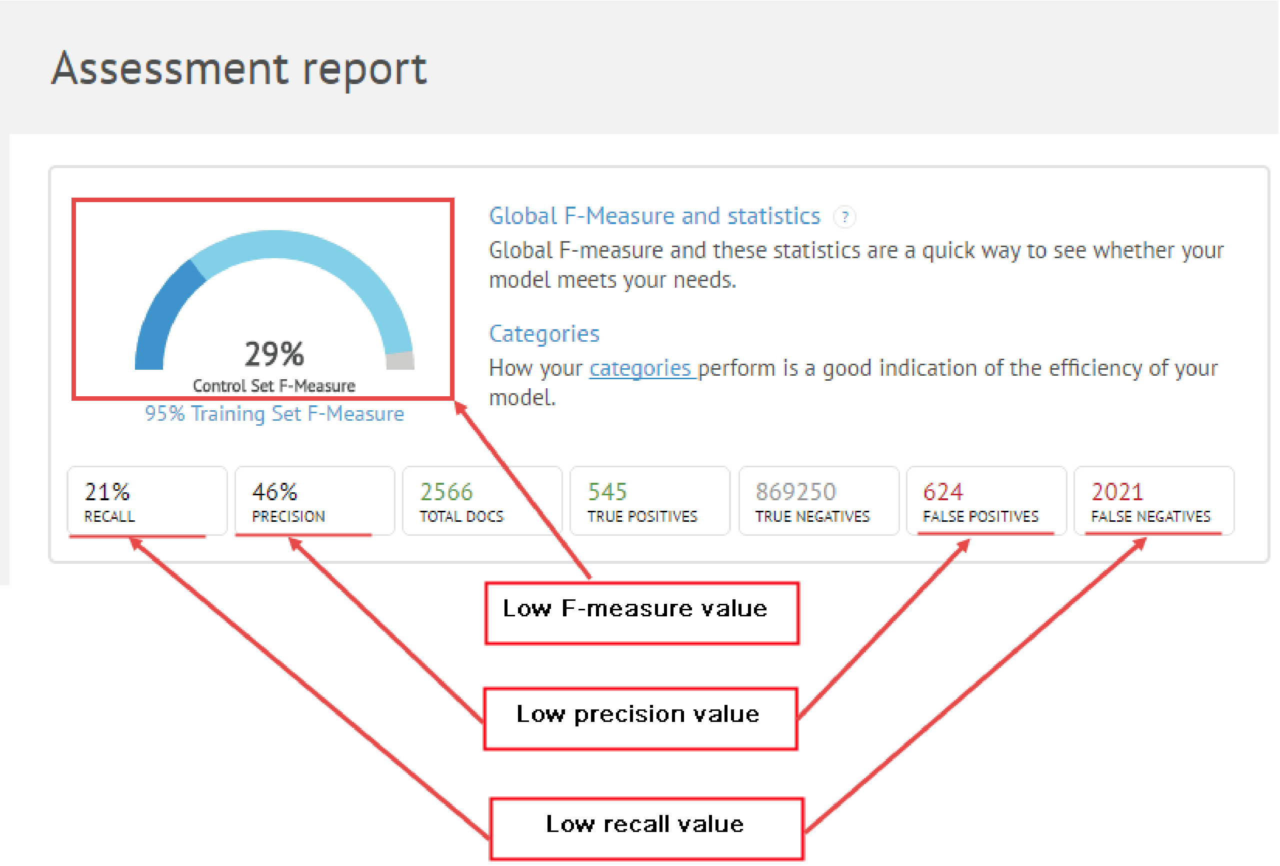
Next, we consider some of the possible reasons why your classification model may fall short of your expectations. The following glossary terms are used:

- FP
- FN
- Precision
- Recall

- F-measure

Improvements at the model level

To review the quality of your model at the model level, open the assessment report by clicking the **Training Set F-Measure** tab or **Control Set F-Measure** tab at the top of the project home page.



Assessment report

29%
Control Set F-Measure
95% Training Set F-Measure

Global F-Measure and statistics
Global F-measure and these statistics are a quick way to see whether your model meets your needs.

Categories
How your categories perform is a good indication of the efficiency of your model.

21%	46%	2566	545	869250	624	2021
RECALL	PRECISION	TOTAL DOCS	TRUE POSITIVES	TRUE NEGATIVES	FALSE POSITIVES	FALSE NEGATIVES

Low F-measure value

Cause

The low F-measure in this case has been caused by a large number of FPs and FNs among the classified documents.

Solution

Since the F-measure is calculated using the TPs, FPs, and FNs for all classified documents, look for possible causes at the category and document levels.

Low precision and recall values

Cause

Increased precision means fewer FPs, which leads to lower recall values because some of the correct answers are filtered out.

Increased recall means fewer FNs, which leads to lower precision values due to a greater number of FPs.

Solution 1

Depending on which is more important in your project, precision or recall, change the position of the **Inclusiveness** slider in the project settings. ABBYY Smart Classifier Model Editor will adjust the model accordingly.

❗ **Note:** *Please note that the inclusiveness setting only affects the choice of the best model at the training stage, when the program selects the model that keeps the cost of error correction to a minimum (i.e. any other model would require more human effort to correct classification errors). The ratio of FPs to FNs in a training/assessment report will not be the same as that selected by means of the inclusiveness slider.*

The possible positions of the **Inclusiveness** slider are: 1/10; 1/6; 1/3; 1/2; 1; 2; 3; 6; 10. The correct position of the slider is determined by the acceptable ratio of the cost of correcting the FNs to the cost of correcting the FPs in your classification scenario. The cost of correcting an error is the manual effort required from the user to correct the FN or FP errors so that the documents are classified correctly. For example, the 1/2 slider setting means that an FP is two times more expensive to correct than an FN.

Consider the following example.

In you are analyzing technical support requests, ABBYY Smart Classifier will generate a list of knowledge base articles that are most likely to contain a solution to the user's problem. An FP error means that the support engineer has to skim through the candidate article to make sure that it does not contain the answer to the user's question. This will take the engineer approximately 15 seconds.

An FN error means that the support engineer will not be presented with a knowledge base article that answer's the user's question. In this case, the engineer will have to find the relevant article himself. This will take the engineer, say, 5 minutes (300 seconds). In this example, the cost of an FN error is 300 seconds and the cost of an FP error is 15 seconds, i.e. FNs are 20 times more expensive to correct than FPs. In ABBYY Smart Classifier Model Editor, the rightmost position of the **Inclusiveness** slider corresponds to a ratio of 10, which assumes that FNs are 10 times more expensive to correct than FPs.

Solution 2

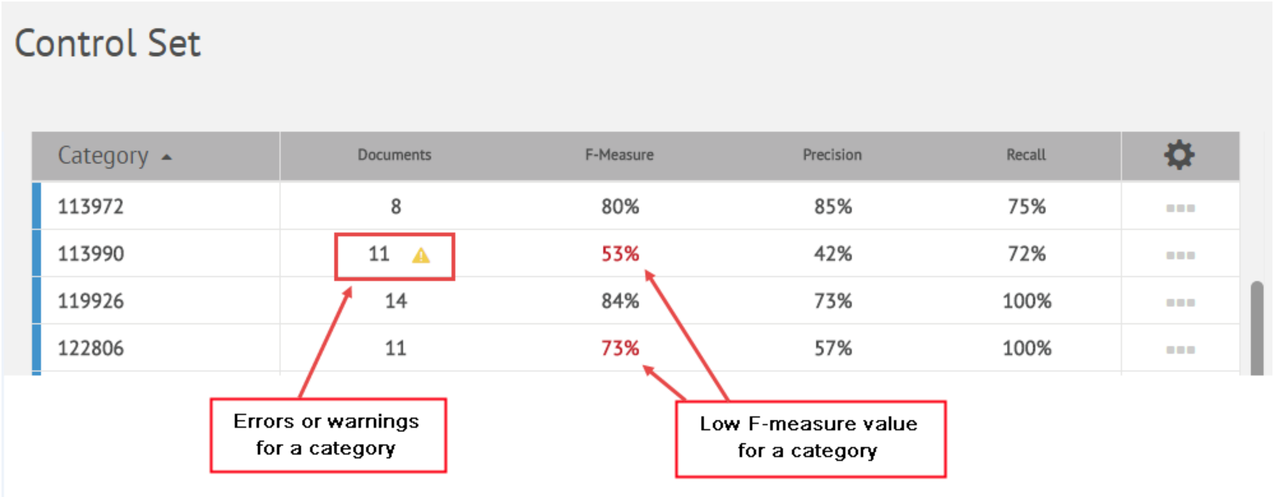
Since the precision and recall values are calculated using the TPs, FPs, and FNs for all classified documents, look for the possible causes at the document level.
To improve precision, you need to reduce the number of FPs. To improve recall, you need to reduce the number of FNs.

Improvements at the category level

When evaluating your model at the category level, look at:

- the F-measure value for each category
- the categories with errors or warnings

To review the quality of your model at the category level, click the **Training Set** tab or **Control Set** tab at the top of the project home page and go to the page listing the categories.



Control Set

Category ▲	Documents	F-Measure	Precision	Recall	⚙
113972	8	80%	85%	75%	⋯
113990	11 ⚠	53%	42%	72%	⋯
119926	14	84%	73%	100%	⋯
122806	11	73%	57%	100%	⋯

Errors or warnings for a category

Low F-measure value for a category

Low F-measure value for a category

Cause 1

The category from the training set contains too few documents.

Solution

Add more relevant documents of this category to the training set.

Adding more documents to the training set will enable ABBYY Smart Classifier to pick out document features more precisely and to optimize its classification algorithm. To build a good quality model, it is recommended to have at least 100 documents in each category.

Cause 2

The category from the control set contains very few documents (10 to 20). If this is the case, any classification errors will have considerable effect on the result.

Suppose a category contains 10 documents, of which one document has been assigned three true categories and the rest have been assigned one true category. Suppose, further, that all but one documents have been classified correctly. The misclassified document has been assigned only one category, which is the same as the true category. In this case, this document has caused 1 TP and 2 FNs. As a result, recall for the category will be 83% (10/(10+2)*100%). Thus, one incorrectly classified document has reduced recall by 17% and the F-measure down to 90% (2*1*0.83/(1+0.83)*100%).

Solution

To have objective metrics for each category, it is recommended to have at least 100 documents per category in a control set.

Cause 3

The text features identified by ABBYY Smart Classifier cause a large number of FNs. To see all documents with FN errors, click the **Show** button above the list of documents of this category and select **False Negatives** from the drop-down list.

Solution

Open the text of the document. The text features having an impact on the classification of this document will be highlighted in orange. If you think that some of the highlighted words (e.g. pronouns, prepositions, etc.) should have no effect on classification, add them to the list of stop words. This will make ABBYY Smart Classifier ignore these words when classifying documents.

Suppose you have a training set made up of news headlines and all of the news in the POLITICS category have the word *April* in their dateline. ABBYY Smart Classifier may well have identified this word as a text feature typical of documents in the POLITICS category. However, documents in the control set containing the word *April* will most likely have nothing to do with politics. To prevent ABBYY Smart Classifier from assigning the POLITICS category to such documents in the control set, you should add the word April to the stop list (you may also want to add all the other months as well).

Important! You must exercise great care when adding words to the stop list, because they will be treated as non-features for all the categories, resulting in improved classification for one category but much worse classification for others. For example, adding the names of the months to the stop list may improve the classification of news items in the POLITICS category, but it may adversely affect the classification of MARKET NEWS, as dates may be an important feature of news in that category. For this reason, it may be preferable to put texts with these words into other categories rather than adding the words to the stop list.

To add words to the stop list:

1. Click .
2. On the **Features and Stop Words** screen, click and save the current stop list to your local hard drive.
3. Open the saved file in any text editor, make any necessary edits, and save the changes (see the Appendix for detailed instructions on creating stop lists).
4. Click .
5. On the **Features and Stop Words** screen, click the **Upload Stop Words** link and load your modified stop list.
6. Click to update your model with the new stop list.

Errors or warnings for a category

Cause

A category contains documents with errors or warnings.

Solution

To view the list of documents for which errors or warnings were reported, click the **Show** button and then select the desired item from the drop-down list.

Open the document, identify the possible causes of the errors or warnings in the text, and make the necessary adjustments.

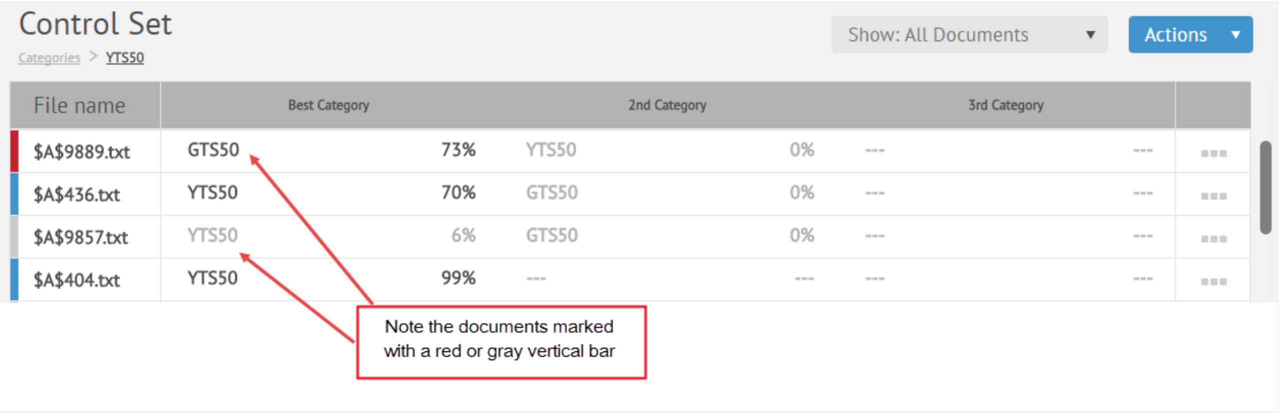
Improvements at the document level

When evaluating your model at the document level, note the documents marked with a red or gray vertical bar.

The documents that have a red vertical bar next to them have been classified incorrectly, i.e. the categories assigned to them are not the same as those manually assigned by an expert.

The documents with a gray bar have no categories assigned to them with a sufficient degree of confidence.

Click a category to open its documents.



	File name	Best Category		2nd Category		3rd Category		
🟥	A9889.txt	GTS50	73%	YTS50	0%	---	---	■■■
🟦	A436.txt	YTS50	70%	GTS50	0%	---	---	■■■
⬜	A9857.txt	YTS50	6%	GTS50	0%	---	---	■■■
🟦	A404.txt	YTS50	99%	---	---	---	---	■■■

Note the documents marked with a red or gray vertical bar

Unclassified or misclassified documents may occur for any of the following reasons:

Cause 1

The expert assigned the wrong category to the document.

Solution

Read the document and view the identified features. If you conclude that a different true category should be assigned to the document, assign the right category by clicking the corresponding button on the page.

Cause 2

The document belongs to several categories, but only one category was assigned by the expert.

Solution

Assign all of the fitting categories to the document.

Cause 3

The training set contains categories that are so close that even a human expert will have difficulty distinguishing between them. This type of error may occur, for example, when classifying knowledge base articles into the categories ERRORS INSTALLING ABBYY SMART CLASSIFIER 2.0 and ERRORS INSTALLING ABBYY SMART CLASSIFIER 2.7 if the version number is not explicitly stated in the articles.

Solution

Merge similar categories into one.

Cause 4

A category contains documents that cover a very broad range of topics (e.g. the PEOPLE category for newswire articles may contain news about crime, politics, finance, etc.). When training the model, ABBYY Smart Classifier will identify too many features, and the category will end up with a large number of redundant documents (FPs).

Solution

Split the broad category into several narrower categories.

Cause 5

There are too few documents in the training set.

Solution

Add more documents to the training set.

Cause 6

Category features were detected in a section of the document that should be ignored for classification purposes.

Solution

Open the document in a text editor and delete the sections that should have no effect on the classification result (e.g. if you are classifying dissertations, keep only the heading, abstract, and introduction). Then delete the original document from the training set and replace it with the modified version.

Less Common Usage Scenarios

ABBYY Smart Classifier can also be used to classify documents in the following less common usage scenarios:

- Classifying multilingual document collections
- Classifying hierarchical document collections

Classifying multilingual document collections

Problem

In the ABBYY Smart Classifier Model Editor, you can select only one language for a project. If you need to classify a document collection made up of documents in multiple different languages, use REST API (see the "Some less common classification scenarios" section of the **ABBYY Compreno Products 2.7 Integration Guide)**.

Classifying hierarchical document collections

Problem

Hierarchical collections with less general categories nested beneath more general ones are currently treated by ABBYY Smart Classifier as if all of the categories were on the same level.

Solution

To make ABBYY Smart Classifier aware of the hierarchy of categories in your collection, we recommend using our sample code designed for processing and managing hierarchical collections (HierarchicalCollectionTrainer + modified SmartClassifierSampleApplication). Please contact the ABBYY technical support service to obtain the sample code

Appendix

Contents:

Glossary

Assigned category is a category assigned to a document using a category assignment method selected by the user.

Candidate category is a category that has been identified by ABBYY Smart Classifier as a possible category for a document. This category will be assigned to the document if it meets the requirements of the selected category assignment method.

Category is a grouping of related documents. Examples of categories include RÉSUMÉ, CONTRACT, FINANCIAL REPORT, etc.

Category assignment method is a parameter that defines which of the candidate categories should be assigned to a document.

Three options are available:

- single candidate category
- top candidate category
- all candidate categories

Category confidence score is a value that indicates the probability of a document belonging to a particular category.

Classification is the process of assigning one or more categories to a document based on its content. For each document, the program calculates category confidence scores. The categories to be used in classification are defined by the user in advance.

Classification model is a set of features typical of each category combined with an algorithm that assigns documents to their appropriate categories. A classification model is created when you train the program to classify documents using a training set.

Classified document is a document to which at least one category has been assigned.

Correctly classified document is a document whose assigned categories are the same as the true categories.

False negative (FN) is incorrect non-assignment of a category to a document. A false negative occurs, for example, if the PERSONAL DATA category has not been assigned to a document that contains personal data.

False positive (FP) is assignment of the wrong category to a document. A false positive occurs, for example, if the SPAM category is assigned to a legitimate e-mail message.

Feature is a characteristic shared by documents of a particular category.

F-measure is a metric that combines precision and recall. An F-measure is a number from 0 to 1 (or 0% to 100%).
The F-measure is calculated using the following formula: $(\beta^2 + 1) * P * R / (\beta^2 * P + R)$, where P is the precision for the category, R is the recall for the category, and β is the inclusiveness setting used in the current model. The largest value the F-measure can have is 1 (when P = R = 1).
The F-measure for a training or control set is calculated using the following formula: $(\beta^2 + 1) * P * R / (\beta^2 * P + R)$, where P is the precision for the training or control set, R is the recall for the training or control set, and β is the inclusiveness setting used in the current model.

Inclusiveness is a parameter which shows which of the two metrics is more important, recall or precision (the user may want to prioritize one over the other depending on his classification needs). Inclusiveness is calculated as a ratio of the cost of correcting a false positive to the cost of correcting a false negative. The cost of correction is the effort required to correct the error and classify the document correctly. By default, inclusiveness is set to 1, which means that the cost of correcting a false positive equals the cost of correcting a false negative.

Misclassified document is a document whose assigned categories are not the same as the true categories.

Precision is a metric used for evaluating classification results and indicating the share of documents that actually belong to a particular category among all the documents to which this category has been assigned correctly. Precision is a number from 0 to 1 (or 0% to 100%).
Precision for a category is calculated as TP/(TP+FP).

Precision for a training or control set of documents is calculated as an arithmetic average across all the categories.

Precision characterizes the system's ability to assign a category only to those documents that actually belong to that category, but it does not take the number of documents into account. The greater the precision for a category, the smaller the number of false positives (i.e. the number of documents to which a category has been assigned incorrectly).

Project contains a training set of documents, a control set of documents, and a trained classification model.

Recall is a metric used for evaluating classification results and indicating the share of documents that actually belong to a particular category among all the documents to which this category has been assigned. Recall is a number from 0 to 1 (or 0% to 100%).
Recall for a category is calculated as TP/(TP+FN).

Recall for a training or control set of documents is calculated as an arithmetic average across all the categories.

Recall characterizes the system's ability to assign a category to documents that belong to that category without overlooking relevant documents.
Recall characterizes the system's ability to assign a category to documents without overlooking relevant documents.
The greater the recall for a category, the smaller the number of false negatives (i.e. fewer relevant documents are missed).

Relevant documents are documents that contain features enabling unambiguous assignment of categories to them.

Training is a process whereby the program establishes a set of features characteristic of each category. A trained classification model is used for classifying documents.

Training set is a set of documents which are positive examples for the categories. For example, the CONTRACT category in a training set may contain commercial contracts, license agreements, etc.

True category is a category that, in the user’s opinion, must be assigned to a particular document in a training or control set of documents as a result of classification.

True negative (TN) is non-assignment of a category to a document that does not belong to that category. A true negative occurs, for example, if the SPAM category has not been assigned to a legitimate e-mail message.

True positive (TP) is assignment of the right category to a document that really belongs to that category. A true positive will occur, for example, if the PERSONAL DATA category is assigned to a document that contains personal data.

Unassigned category is a category that has not been assigned to a document.

Unclassified document is a document to which no categories have been assigned.

Creating stop lists

You can use any simple text editor (e.g. Notepad) to compile and edit lists of stop words. We recommend saving stop word files in Unicode or UTF-8 with BOM. Stop words must be separated with a semi-colon.

A stop list must include the following three sections:

- Exact matches.

  This section contains words that should not be used as classification features if they occur in a text exactly as spelled in this section. Note, however, that capitalization will be ignored and *computer* and *Computer*, for example, will be treated as one and the same word.

- Inflected forms.

  This section contains words in canonical dictionary form that should not be used as classification features if they occur in a text in any form, inflected or otherwise. Note, however, that capitalization will be ignored. For example, if you include the operator in this section, the program will ignore *operator*, *operators*, *operator's*, *operators'*, *Operators*, etc.

- Regular expressions.

  This section contains POSIX regular expressions that specify templates for words that should not be used as classification features. Note that regular expressions are case-sensitive and [^A] and [^a] are templates for two different words.

The most commonly used special symbols are:

.	Matches any single character.
?	Matches the preceding character zero times or once.
*	Matches the preceding character zero or more times.
+	Matches the preceding character one or more times.
[]	Encloses a character set any of which can be matched.
|	OR.
[^]	Matches a single character that is not contained within the brackets.

Important! If any of the above sections should be left empty, insert a blank line for the empty section.

Example 1

```
April; May; Jul; e.g.

director
```

```
[^a-z].*
```

If this list of stop words is used, the following words will NOT be used as classification features:

- The words *April, May, Jul, e.g., E.g*.
- Any inflected forms of director and the form director itself: *director, directors, Director's,* etc.
- Any words that begin with any character other than an English lower-case letter, e.g. *Contract, 1*.

Example 2 (no "Inflected forms" section)

```
April; May; Jul; e.g.

[^A-Z].*
```

A blank line is inserted instead of the "Inflected forms" section.

If this list of stop words is used, the following words will NOT be used as classification features:

- The words *April, May, Jul, e.g., E.g*.
- Any words that begin with any character other than an English upper-case letter, e.g. *contract*, *25*.

Supported formats

ABBYY Smart Classifier can automatically classify incoming documents in any of the following formats:

- RTF documents (*.rtf)
- Microsoft Word 97-2003 documents (*.doc)
- Microsoft Word documents (*.docx)
- Microsoft Word macro-enabled documents (*.docm)
- Microsoft Word XML documents (*.xml)
- Unformatted text documents (*.txt) (We recommend saving text files in Unicode or UTF-8 with BOM)
- Web pages (*.html, *.htm)
- Microsoft PowerPoint 97-2003 presentations (*.ppt, *.pps)
- Microsoft PowerPoint presentations (*.pptx, *.ppsx)
- Microsoft PowerPoint macro-enabled presentations (*.pptm, *.ppsm)
- Microsoft PowerPoint XML presentations (*.xml)
- Microsoft Excel 97-2003 workbooks (*.xls)
- Microsoft Excel workbooks (*.xlsx)
- Microsoft Excel macro-enabled workbooks (*.xlsm)
- Adobe InDesign Markup (IDML) documents (*.idml)
- OpenDocument texts (*.odt)
- OpenDocument presentations (*.odp)
- OpenDocument spreadsheets (*.ods)
- Adobe FrameMaker documents (*.mif)
- Adobe PDF documents (*.pdf) (license required)
- Image files (*.jpeg, *.jpg, *.bmp, *.gif, *.tif, *.tiff, *.png, *.djvu, *.dcx, *.dib, *.jb2, *.jp2, *.j2k, *.jpf, *.jpx, *.pcx, *.wdp) (license required)

  Important! *Support for ***.djvu** will be discontinued in future versions of ABBYY Smart Classifier. Please contact ABBYY if you need to process files in this format (see the Technical support section for contact details).*

Supported languages

ABBYY Smart Classifier can classify documents in any of the following languages. Your license determines which of these languages are available to you.

- Armenian
- Azeri (Latin)
- Bashkir
- Bulgarian
- Catalan
- Chinese (Simplified)
- Chinese (Traditional)
- Croatian
- Czech
- Danish
- Dutch
- English
- Estonian
- Finnish
- French
- German
- Greek
- Hungarian
- Indonesian
- Italian
- Japanese
- Kazakh
- Korean
- Latvian
- Lithuanian
- Norwegian (Bokmål)
- Norwegian (Nynorsk)
- Polish
- Portuguese (Brazil)
- Portuguese (Standard)
- Romanian
- Russian
- Slovak

- Slovenian
- Spanish
- Swedish
- Tatar
- Turkish
- Ukrainian

Technical Support

Should you encounter any problems using the program, please first consult the accompanying documentation. You may also want to contact the IT specialist responsible for the correct operation of the software used by your company.

If the problem persists, please ask your IT specialist to contact the ABBYY technical support service via this form on our website:

http://go.abbyy.com/?target=onlinesupport&product=SmartClassifier&lang=en

Our technical support engineers will need the following information in order to address your problem:

- Your full name
- The name of your company
- Your phone, fax, or e-mail address
- The serial number of your license
- The build number of your copy of the product
- A general description of the problem and the complete text of the error message (if any)
- The version of your operating system
- Any other information that you think is relevant