Veritas eDiscovery Platform™

Transparent Predictive Coding User Guide

10.1

VERITAS™

Veritas eDiscovery Platform™: Transparent Predictive Coding User Guide

The software described in this book is furnished under a license agreement and may be used only in accordance with the terms of the agreement.

Last updated: 2021-11-20.

Legal Notice

Veritas Technologies LLC
2625 Augustine Drive
Santa Clara, CA 95054
http://www.veritas.com

Contents

Transparent Predictive Coding User Guide

This guide provides case administrators of the Veritas eDiscovery Platform with details on how to model, prepare, analyze, and review case data using the predictive coding feature.

This section contains the following sections:

About This Guide

As a supplement for case administrators, and a companion to the Case Administration Guide, the Transparent Predictive Coding User Guide is intended to guide you through the steps of applying and monitoring the Transparent Predictive coding feature. It describes the best practices on how to train the system to quickly cull through case data for review, and how to review and tag those items as "trained" content.

Note: You must have the Case Administrator role, or have appropriate administrator permissions to perform many of the tasks associated with the Predictive Coding feature. Refer also to the *Case Administration Guide* for more information.

Revision History

The following table lists the information that has been revised or added since the initial release of this document. The table also lists the revision date for these changes.

Revision Date	New Information
December 2021	• Updated version for release 10.1
March 2021	• Minor edits
October 2018	• Minor edits
June 2017	• Minor edits
July 2016	• Branding and minor edits
August 2015	• Remove Rights Management Guide
March 2015	• Image accessibility • Branding and minor edits
October 2014	• Updated screenshots • Branding edits

Revision Date	New Information
November 2013	• Modified workflow from optional to required Additional Sample step • Updated Controlled Prediction Accuracy Report • Updated Prediction Accuracy Report • Added Advanced Search Predictions option • Added Prediction Status ability to view items for Initial and Additional Sample.
June 2013	• Addition of Prediction Test Report • Addition of Controlled Prediction Test Report • Addition of Prediction Status Graph • General updates to graphics, text and minor edits
Sept 2012	• New guide—Documents new predictive coding feature, including import/export modeling workflow, and review, tagging, and filtering capabilities within the Analysis & Review module including: – separate tagging of attachments (refer also to the *User Guide*) – search results filtering and review enhancements for attachments – searchable ID for attachments and embeddings – search report and tag event history search enhancements

Technical Support

Technical Support maintains support centers globally. All support services will be delivered in accordance with your support agreement and the then-current enterprise technical support policies.

For information about our support offerings and how to contact Technical Support, visit our website:

https://www.veritas.com/support

You can manage your Veritas account information at the following URL:

https://my.veritas.com

If you have questions regarding an existing support agreement, please email the support agreement administration team for your region as follows:

Worldwide (except Japan)	CustomerCare@veritas.com
Japan	CustomerCare_Japan@veritas.com

Documentation

Make sure that you have the current version of the documentation. The latest documentation is available from:

- **Documentation** link at the bottom of any page in the Veritas eDiscovery Platform landing page.
- **Veritas Technologies LLC. Products Web site:** https://www.veritas.com/product/a-to-z

Documentation Feedback

Your feedback is important to us. Suggest improvements or report errors or omissions to the documentation. Include the document title, document version, chapter title, and section title of the text on which you are reporting. Send feedback to:

eDiscovery.InfoDev@veritas.com

You can also see documentation information or ask a question on the Veritas community site.

https://vox.veritas.com/

Overview

Release 7.1.2 of the Veritas eDiscovery Platform adds Transparent Predictive Coding to its technology-assisted review offerings. Transparent Predictive Coding helps organizations defensibly reduce the time and cost of document review. Transparent Predictive Coding enables legal teams to automatically classify documents according to case issues such as responsiveness or privilege, often requiring only a small fraction of the case to be reviewed manually. The net result is a more defensible automated document review process at significantly reduced cost.

This chapter contains the following topics:,

What is Transparent Predictive Coding?

Over the last several decades there have been significant advances in the area of computing called machine learning. The goal of this research is to improve the ability of software to learn from inputs in the environment and use this information to make decisions. This machine learning technology can also be used to help make the review process more cost effective and accurate. For eDiscovery, this is commonly referred to as *predictive coding* (other terms are suggestive coding, computer-categorized document review and predictive categorization).

Veritas's Transparent Predictive Coding technology provides visibility into the prediction process, enabling more informed decisions and facilitating greater review accuracy. This solution provides a workflow that adapts to the unique requirements of each case, allowing reviewers to begin using predictive coding immediately and achieve optimal results. Each step in the workflow is documented by comprehensive reporting to help demonstrate the integrity of review to the court.

How Does Predictive Coding Work?

Predictive coding works by having software interact with human reviewers in order to learn the review criteria in the case. As reviewers tag documents in a sample set, the software learns the criteria for assessing documents and can generate accurate estimates of which tags should be applied to the remaining documents. As a result, fewer documents need to be reviewed and reviewers can tag documents more efficiently thereby reducing review costs.

What Predictive Coding is Not

Predictive coding technology is often confused with other types of technology assisted review tools like concept search and clustering. However, unlike those technologies which automatically extract relationships and patterns from documents without human intervention, predictive coding requires human input from the outset. Predictive coding involves training and fine tuning the computer to accurately classify documents as, for example, responsive and not responsive. This requirement is because predictive coding goes a step further than other types of tools to make it easier to search, group, or organize documents.

Using Transparent Predictive Coding for Optimal Performance

Transparent predictive coding is best suited for cases and situations where there is a large amount of documents to review. This means that small review efforts (fewer than 100,000 documents or less than 20 GB) might be better suited for other review strategies and workflows.

The system works best on extracted text documents. Currently, non-text information such as images that cannot be OCR processed, sound, and video are not optimal candidates for transparent predictive coding.

Types of Transparent Predictive Coding Users

In Transparent Predictive Coding, the user type controls what operations users can perform on the Transparent Predictive Coding model and system. The various tasks and actions in the Transparent Predictive Coding workflow are divided between two classes or types of user. Each type of user has different responsibilities and access to different levels of Transparent Predictive Coding features and tasks.

- **Case Administrators and Case Managers** — have full control over the Transparent Predictive Coding system. They can create, modify, monitor and manage the end-to-end Transparent Predictive Coding workflow and model. Additionally, they can set access profiles for reviewers and other key case team members that enable them to perform a subset of tasks and views.
- **Reviewers, Case Users, and Key Review Team Members** — can search on predictions, view, sort, and filter prediction ranks and perform tasks related to review analysis. Reviewers in this class are restricted from making system wide changes to the Transparent Predictive Coding model and workflow.

Below is a summary of Transparent Predictive Coding tasks and features:

	Case Administrators Case Managers	**Reviewers/Case Users Key Review Team Members**
Actions	✓	
• Controlled Prediction Accuracy Test	✓	
• Train	✓	
• Next Set	✓	
• Predict	✓	
• Prediction Test	✓	
Transparent Predictive Coding Management	✓	
• Prediction Status	✓	
• Export Predictive Tag	✓	
• Access Profile Changes	✓	
Review Dashboard	✓	✓
• Prediction Ranks	✓	✓
Search	✓	✓
• Prediction Ranks	✓	✓

Transparent Predictive Coding Best Practices

This section describes what you should know and consider for successfully creating, building and tuning your Predictive Tag.

Refer to the following topics in this section:

Understanding the Workflow

Transparent Predictive Coding utilizes a loosely-coupled workflow to ensure full flexibility in its usage. It is uniquely designed to complement existing review workflows. The following information should be taken into account before identifying the set of documents for which Transparent Predictive Coding is to be applied.

Items with No Text

The text content of items is used for both training the system and assigning Prediction Ranks. In order to achieve the highest accuracy with Transparent Predictive Coding, items that have no text content should be separated for manual review and not put through the Transparent Predictive Coding workflow. In addition, any items with processing errors should also be separated for manual review.

Early Case Assessment

Transparent Predictive Coding has powerful analytics that enable effective Early Case Assessment. These analytics can accurately identify and separate items which are determined to be unrelated to the case.

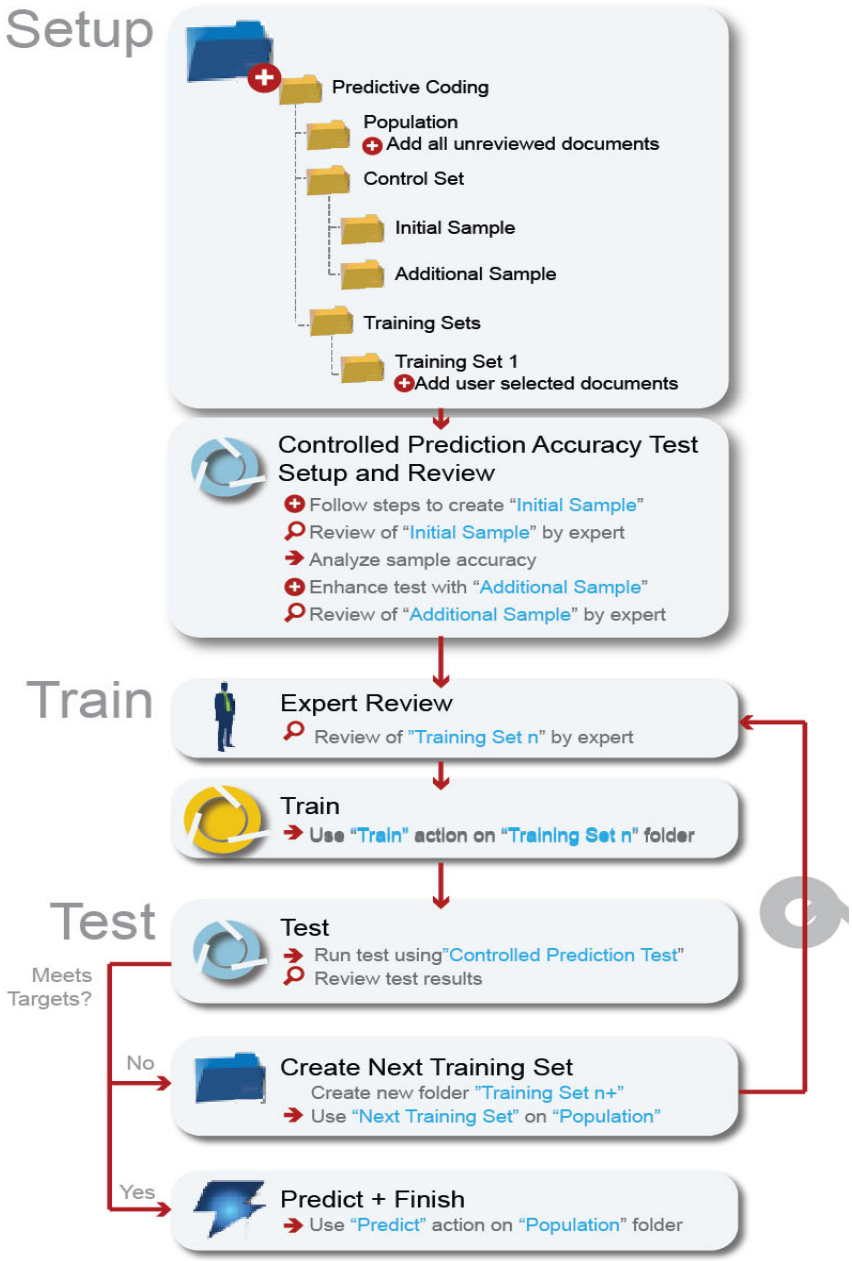
Transparent Predictive Coding Phases

Review workflows typically involve using multiple tags across various tag sets. To achieve the highest accuracy across multiple tags in the most efficient way possible, the training and testing of individual tags should be phased as much as possible.

Best Practices Workflow

The following workflow diagram describes the best practices for identifying all of the positive items in a large set and utilizing the most powerful features in Transparent Predictive Coding. This workflow leverages the system's statistically valid sampling and ensures the richest possible measures of accuracy.

Note: The diagram details of Setup, Train and Test are covered in *"Workflow Walkthrough" on page 17*.



Folders

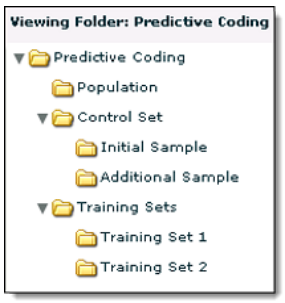
Manually create the following folders to aid the management of the recommended workflow.

Table 1: Folder Setup

Folder	User or System Populated	Description
Predictive Coding		Parent folder containing the documents involved in the Transparent Predictive Coding workflow.
Population	User	Subfolder of Predictive Coding containing unreviewed documents in the Transparent Predictive Coding workflow.
Control Set		Subfolder of Predictive Coding containing the documents selected by the Controlled Prediction Accuracy Test action for use in evaluating the accuracy of the Predictive Tag.
Initial Sample	System	Subfolder of Control Set containing the documents selected by Step 1 of the Controlled Prediction Accuracy Test action.
Additional Sample	System	Subfolder of the Control Set containing the documents selected by Step 2 of the Controlled Prediction Accuracy Test action
Training Sets		Subfolder of Predictive Coding containing documents to be used by the Train action for training the system
Training Set 1	User	Subfolder of Training Sets containing a user selected set of documents to be used for the initial training of the system. The contents of this folder should be carefully selected from the Population by using the system's search features to locate between 100-1,000 items for initial review. For best results, this set of items should have about 50% positive items and 50% negative items.
Training Set 2 - N	System	Subfolders of Training Sets containing the documents selected by the Next Training Set action to be used for iterative training of the system.

Folder Structure

The following Predictive Coding folder hierarchy follows the suggested folder setup guidelines.



- At the top level is the main folder: Predictive Coding
- Next level are three folders: Population, Control Set and Training Sets
- Control Set has a minimum of two subfolders: Initial Sample and Additional Samples
- Training Set has subfolders that correspond to each training set: Training Set 1 and Training Set2.

Workflow Walkthrough

Reference

The following is a reference for the recommended workflow.

Table 2: Workflow Reference Guide

Step	Sub-Step
Setup	1. Setup initial folder structure 2. Use **Controlled Prediction Accuracy Test** to create the Initial Sample for the Control Set 3. Review the Initial Sample to calculate the yield 4. Use **Controlled Prediction Accuracy Test** to create the Additional Sample for the Control Set 5. Review the Additional Sample to finish the Control Set
Iterate	1. Select items from the Population to become Training Set 1 2. Review all unreviewed items in Training Sets 3. Use **Train** to Train on all items in Training Sets 4. Use **Controlled Prediction Accuracy Test** to Predict and Test on all items in Control Set 5. If results are unsatisfactory, use **Next Training Set** to have the system select the next training items 6. Repeat Steps 2-5 until results are satisfactory
Finish	1. Use **Predict** to Predict on the Population 2. Use the **Prediction Rank Threshold** from the last Test to move positive items out of Transparent Predictive Coding

Detailed Workflow

The following sections describe the individual steps in this workflow. For details on each of the Transparent Predictive Coding actions, see *"Transparent Predictive Coding Actions" on page 25*.

Setup

The purpose of this section is to help you set up a set of folders for use in Transparent Predictive Coding.

SETUP STEPS

A. Setup initial folder structure

1. Create the Predictive Coding, Control Set, Initial Sample, Additional Sample, Population, Training Sets, and Training Set 1 folder (see *"Folder Setup" on page 15*)
2. Place all items to go through Transparent Predictive Coding into the Population folder
3. Select items for placement into the Training Set 1 folder
   A. Navigate to the Population folder
   B. Use the search features to locate between 100-1,000 items. See *"Training Set 1" on page 15* for more details.
   C. Use the Folder action to Move these items to the Training Set 1 folder

B. Use Controlled Prediction Accuracy Test to create the Initial Sample for the Control Set

1. Open the Controlled Prediction Accuracy Test action
2. Select the Tag to put through Transparent Predictive Coding
3. Under Step 1, specify the Population folder as the Population Source
4. Specify the Confidence level
5. Specify the Margin of Error
6. Specify the Initial Sample as the Destination Folder
7. Finish and generate the Initial Sample

C. Review the Initial Sample to calculate the yields

1. Navigate to the Initial Sample folder
2. Review all items in the folder

SETUP STEPS

D. Use Controlled Prediction Accuracy Test to create the Additional Sample for the Control Set

1. Open the Controlled Prediction Accuracy Test action
2. Under Step 2, check the box "The initial sample has been reviewed."
3. Specify the Target F-Measure
4. Specify the Target F-Measure Range
5. Specify the Additional Sample as the Destination Folder
6. Finish and generate the Additional Sample

E. Review the Additional Sample to finish the Control Set

1. Navigate to the Additional Sample folder
2. Review all items in the folder

Train and Test

The purpose of this step is to achieve the highest quality Predictive Tag.

TRAIN & TEST STEPS

A. Review all unreviewed items in Training Set

1. Navigate to the Training Set 1 folder (where Training Set 1 is the current iteration)
2. Review all items in the folder

B. Use Train action on all items in Training Set

1. Open the Train action
2. Select the Predictive Tag that is going through Transparent Predictive Coding
3. Select "All items in search results" and ensure that "Include previously trained items" is checked to select all items in Training Sets for training

C. Use Controlled Prediction Accuracy Test to Predict and Test on all items in Control Set

1. Open the Controlled Prediction Accuracy Test action
2. Under Step 3, ensure that "The final sample has been reviewed." is checked
3. Finish and run the Controlled Prediction Accuracy Test
4. When the Controlled Prediction Accuracy Test job is complete, open the Test Report and evaluate the results

D. If results are unsatisfactory, use Next Training Set to have the system select the next training items

1. Navigate to the Population folder
2. Open the Next Training Set action
3. Select the Predictive Tag that is going through Transparent Predictive Coding
4. Select "All items in search results" to select from all items in the Population to be considered for the next training set
5. Specify a Move action to a new Training Set 2 folder
6. Finish and generate the Training Set 2 folder

E. Repeat Steps B through E until results are satisfactory

Predict and Finish

The purpose of this step is to identify positive items to move out of Transparent Predictive Coding.

PREDICT & FINISH STEPS
A. Use Predict to Predict on the Population
1. Navigate to the Population folder
2. Open the Predict action
3. Select the Predictive Tag that is going through Transparent Predictive Coding
4. Select "All items in search results" to assign Prediction Ranks to all items in the Population
5. Finish and Predict on the entire Population
B. Use the Prediction Rank Threshold from the last Test to move positive items out of Transparent Predictive Coding
1. From the last Test Report, get the Best Prediction Rank Threshold
2. Navigate to the Predictive Coding folder
3. Run an Advanced Search for all items at and above the Best Prediction Rank Threshold
4. Use either the Tag or Folder action to mark these items as predicted to be positive

Alternate Workflows

Transparent Predictive Coding is a highly flexible feature that be integrated into a number of alternate workflows. The following sections describe two potential workflows that use elements of the Transparent Predictive Coding features.

Prioritizing Linear Review

To prioritize a linear review, simply select a Training Set to review, use the Train action on the items to Train the system, and then use the Predict action on all of the items in your case. All items will now have Prediction Ranks, which you may sort on in descending order for prioritization.

Using Alternate Sampling Methods

To utilize alternate sampling methods for your workflow, simply create a sample folder manually with the Folder action. Follow the same workflow as described in the Best Practices Workflow, but instead of using the Controlled Prediction Accuracy Test action, use the Predict action on your sample folder and the Prediction Accuracy Test action on your sample folder to generate a test report.

Note: The Prediction Accuracy Test generates results on the set of items fed into the action. Unlike the Controlled Prediction Accuracy Test, it does not project any results onto the Population.

Using Transparent Predictive Coding

This section takes you through the steps for building and tuning a Transparent Predictive Coding review system. You create, simulate, and build the Transparent Predictive Coding review intelligence with input from your expert reviewers and the Transparent Predictive Coding system. It is an iterative process.

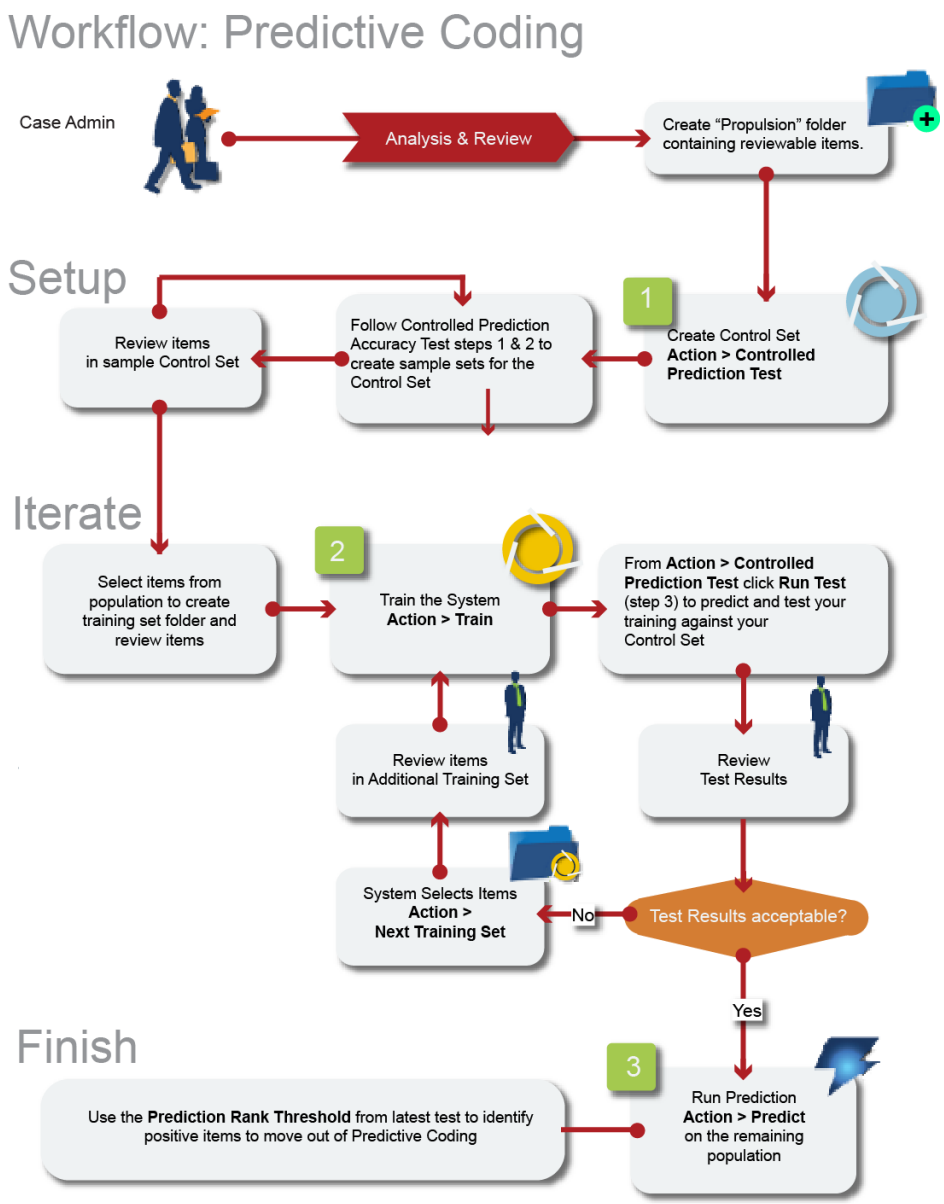
The goal is to have the Transparent Predictive Coding system learn review criteria for classifying items as responsive, privileged, or any other issue code through input from expert reviews. Essentially, you are training and teaching the system through review control and training sets. Over time, the Transparent Predictive Coding system becomes statistically stable, learning enough to correctly predict and rank training sets of documents, progressing to a point where it can be leveraged against your entire item population, and enabling you to achieve high levels of review accuracy at significantly reduced time and cost.

Refer to the following topics in this section:

Workflow and Operating Considerations

The system's flexibility allows you to follow different workflows depending on your review and case requirements. For example, you may perform only some of the Transparent Predictive Coding steps, or perform the steps by varying the sequence, depending on the case review requirements and workflow. As an iterative process, some steps or stages in the system may be repeated a number of times to refine the approach after a better understanding of the data emerges. For specific workflow details, see *"Transparent Predictive Coding Best Practices" on page 13*.

Note: All actions, predictions, and tags operate on the item-level.



Transparent Predictive Coding Actions

The Transparent Predictive Coding workflow combines the efforts of your expert reviewers with the following system actions:

Table 3: Transparent Predictive Coding Actions

Icon	Action	Description
	Controlled Prediction Test	Setup and run a statistically valid test of prediction accuracy.
	Train	Supply reviewed items to system for training
	Next Training Set	Get recommended set of items to review to increase accuracy.
	Predict	Apply prediction ranks to items based on system knowledge.
	Prediction Test	Run independent testing on any set of items.

Summary chart of the Transparent Predictive Coding action workflow.

Table 4: High-Level Summary of Action Workflow

Actions	Method	Inputs	Usage Guidelines - Dependencies
Controlled Prediction Accuracy Test **Goal:** Create an accurate control set	Switch between human review and Transparent Predictive Coding calculator and guided steps 1-3 May need to repeat steps until you reach desired control set accuracy.	Population Single tag	Prerequisites: Manually create folder hierarchy and store population data **Caution:** Do not add or remove removing items from the control set. If you do, you must re-create and re-review the control set.
Train **Goal:** Identify a sufficient training sample	Follow the Training steps	Reviewed and tagged training set	Prerequisites: Training items must be viewable within current search. **Caution:** Training conflict errors should be resolved. If not, it can potentially lower the prediction accuracy.
Predict **Goal:** Arrive at an acceptable prediction ranking	Follow the Predict steps	Control set Population	**Caution:** Do not submit training sets to the Predict action.
Next Training Set **Goal**: Fetch the next set of items for train	Follow the Next Training Set steps	Population Predictive tag	Prerequisites: Population items must be viewable within current search or in a designated folder.
Prediction Accuracy Test **Goal:** Quickly test current learning state against random set of reviewed items.	Follow the Prediction Accuracy Test steps	Control Set *or* Items that have completed the Predict action.	Prerequisites: Items must be viewable within current search

Creating a Control Set

The action Controlled Prediction Accuracy Test produces a highly reliable sample which is then used to test prediction rank accuracy. The goal of Controlled Prediction Accuracy Test is to create an acceptable control set that can be further leveraged against training sets. The workflow involves switching between the work product from your expert reviewers and the Transparent Predictive Coding calculator in order to evaluate and finalize the quality of the control set.

	Important:	Do not delete initial sample or initial sample folders.
	Prerequisites:	A folder hierarchy, set up manually, to organize data. For help on creating folders, see *"Folders" on page 15*
	Inputs:	Population data exists and resides in the folder structure A single tag
	User:	Case Administrator and Case Managers

Selecting a Controlled Prediction Accuracy Test Tag

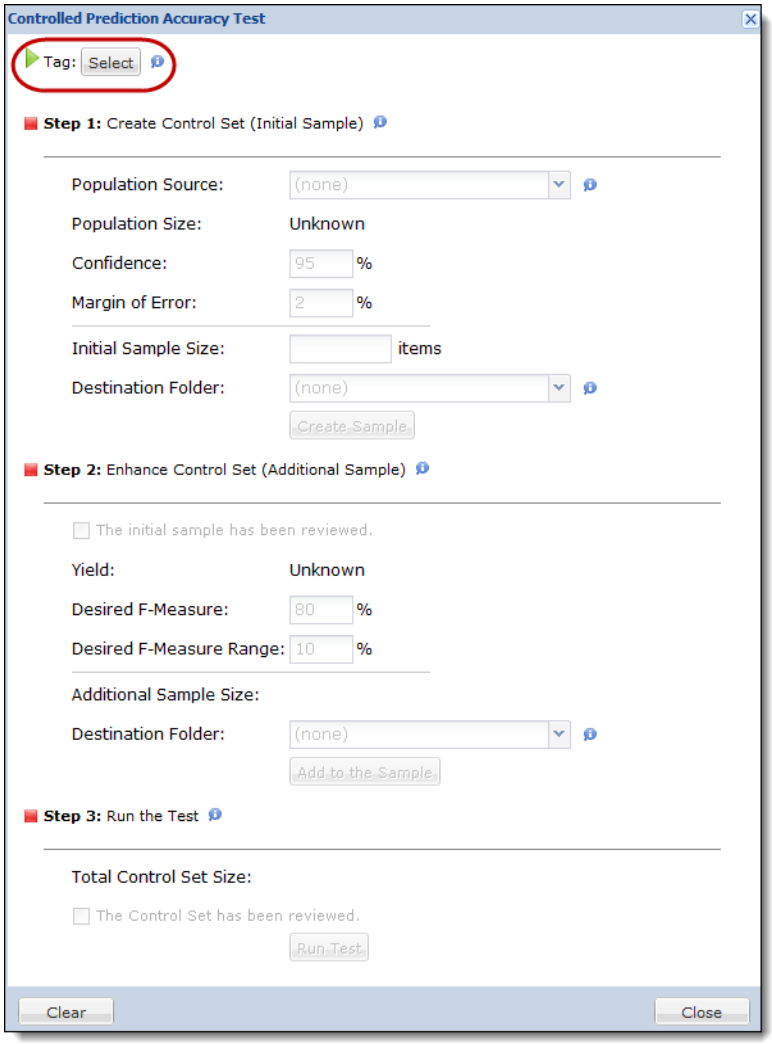
The first task is to select a predictive tag that you want to train the system with.

To select a Controlled Prediction Accuracy Test Tag

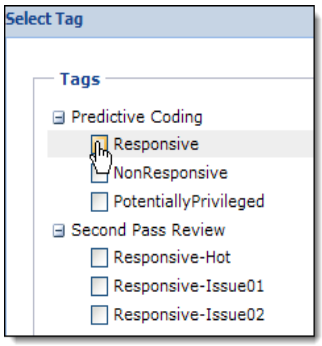
1. From the Documents Screen under the Analysis & Review module, click **Action** and select **Controlled Prediction Accuracy Test** to start training the system.

   The Controlled Prediction Accuracy Test menu displays.

2. From the Tag section at the top of the Controlled Prediction Accuracy Test screen, click **Select**.



The **Select Tag** screen opens.



3. Select the tag for the control set.

   Note: Each tag should have its own control set. This is because each tag is associated with a different accuracy and should be measured and validated separately.

Step 1: Create Control Set (Initial Sample)

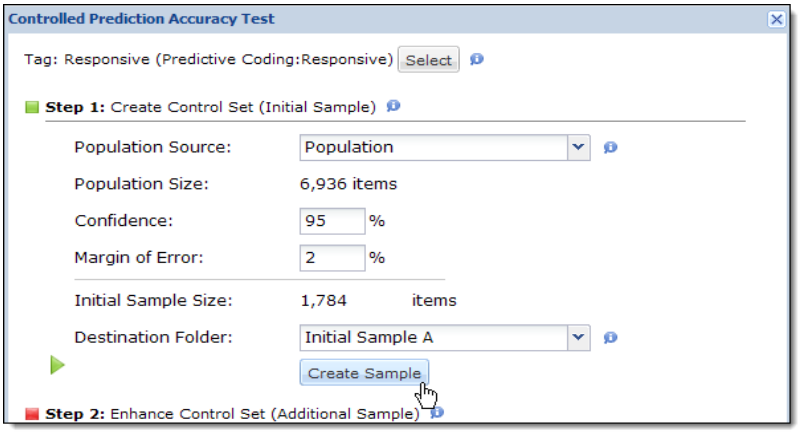
Create the first statistically valid sample using the measurements below. Return to this screen after reviewing this sample.

1. Specify the following Create Sample field information:

Table 5: Controlled Prediction Test - Step 1: Create Sample

Field	Description
Population Source	Select the source that contains only the set of items that you want to test against or sample from.
Population Size	Automatically displays the number of items in the Population Source.
Confidence	Specify a confidence level which is a statistical measure of certainty expressed as a percentage. Confidence level indicates the likelihood that the actual number is within the margin of error. For instance, if the confidence level is 95% with a margin of error of ±3%, then there is a 5% chance the actual value lies outside the margin of error. The default is 95%.
Margin of Error	Specify the desired margin of error. Margin of error is a statistic expressing the amount of random sampling error related to an estimate. For example, consider a situation where a random sampling of voters is polled to reach an estimate that 60% of voters intend to vote for candidate A in the next presidential election. If the margin of error for the estimate is ±3%, then somewhere between 57-63 percent of the voters are estimated to vote for candidate A. Importantly, all things being equal, a larger sample size will produce a smaller margin of error.
Initial Sample Size	Automatically displays the minimum number of items needed to achieve the desired confidence level and margin of error.
Destination Folder	Specify the folder where the control sample of documents will be moved to or copied. As part of the Best Practices recommendation, see *"Folders" on page 15* (for example, Control Set: Initial Sample). **Important:** The folder you specify must be empty.

2. After filling in all the initial sample fields, click **Create Sample** to launch the sample creation job.



Note: The actual folder size may be larger than the sample size. This is because the items sampled may be part of larger document families.

3. Once you have created an initial sample, have your expert reviewers review it before continuing with the next step.

   There are two ways to view the items selected for the Initial Sample:

   – From the Prediction Status page, for the relevant job history row, click on the magnifying glass icon under the Actions column to access the results.

   – From the Advanced Search page, Predictions section, select **Only Initial Sample**.

Step 2: Create Additional Sample

Increase the quality of your statistically valid sample using the additional measurements listed in this section. Return to this screen after reviewing your sample.

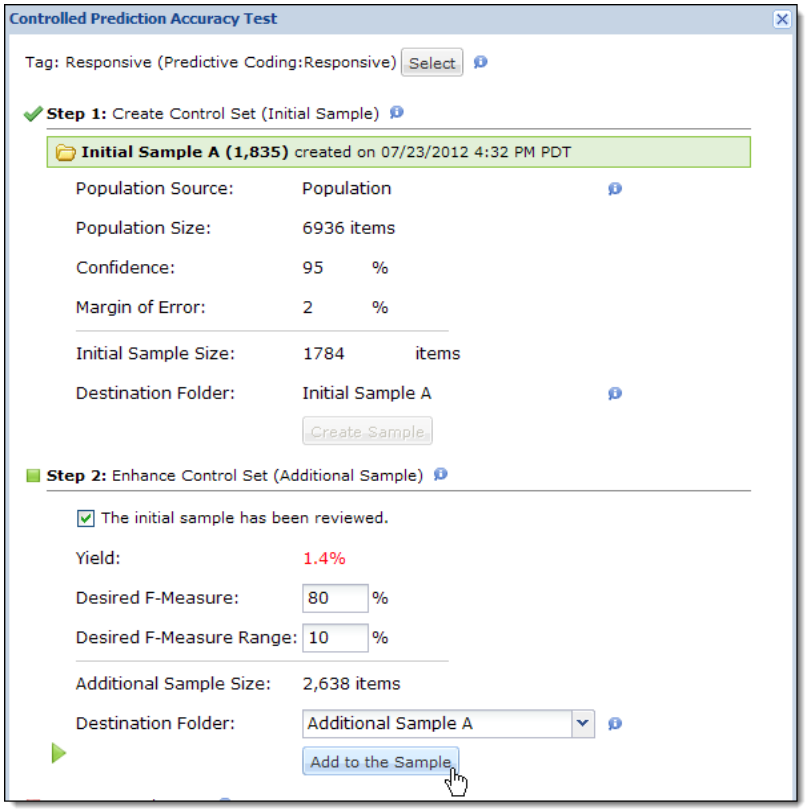
1. Check that **The initial sample has been reviewed** by your expert reviewers.

2. Specify the following field information

Table 6: Controlled Prediction Test- Step 2: Additional Sample

Field	Description
Yield	Automatically displays the computed yield from your review team's sample. The yield is computed on the initial sample. Yield is a measure based on the amount of items tagged with your Predictive Tag which is the percentage of documents in the population which are responsive. A low yield often causes the margin of error to increase, so additional sampling may be required. **Sample Errors to Avoid:** • **0%**—The sample was probably not reviewed. • **100%**— The sample size contains only positive items and the tag is not suitable for use by the system.
Desired F-Measure	Measures the accuracy on a test sample. The values must be between 1 and 100 and the default is 80. **Note:** A higher Desired F-Measure results in a smaller Additional Sample Size.
Desired F-Measure Range	The range of acceptable F-Measure values. The default is 10. For example, if the F-Measure is 80 and the Desired F-Measure Range is 10, then the results will be between 75 and 85. **Note:** A higher Desired F-Measure Range results in a smaller Additional Sample Size. The Desired F-Measure is guaranteed once the Desired F-Measure is reached. This may be a useful measurement for defensibility. A low range means that more items will need to be reviewed.
Additional Sample Size	Automatically displays the statistically valid sample size required based on Yield, Desired F-Measure, and Desired F-Measure Range calculations.
Destination Folder	Specify a folder for the sample. As part of the Best Practices recommendation, see *"Folders" on page 15* (for example, Control Set: Initial Sample). **Important:** The folder you specify must be empty.

3. After supplying all the required Additional Sample fields, including a destination folder (which must be empty), click **Add to the Sample**.



Note: The actual folder size may be larger than the sample size. This is because the items sampled may be part of larger document families.

4. Once you have created an additional sample, have your expert reviewers review it before continuing with the next step.

   There are two ways to view the items selected for the Additional Sample:

   – From the Prediction Status page, for the relevant job history row, click on the magnifying glass icon under the Actions column to access the results.

   – From the Advanced Search page, Predictions section, select **Only Additional Sample**.

Step 3: Run the Test

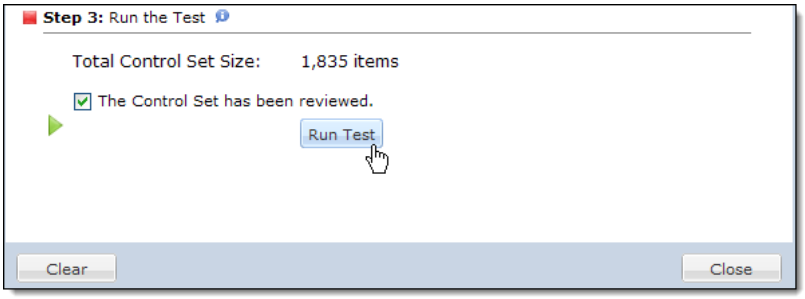
Before you proceed, Step 3 assumes that you have selected items for the initial and additional samples (Steps 1 and 2). During Step 3, the system compares the human tagging decisions against its prediction ranks and provides a report. The report gives a ranking of correct and incorrect predictions. The rankings in the report give you an idea of what to expect if you were to bulk tag the remaining items.

Important: You must run Train on the selected tag before running the Controlled Prediction Accuracy Test.

To run the Controlled Prediction Accuracy Test on the sample set

1. Verify that the **Total Control Set Size** displays. This should be a cumulative count of items in both sample folders from Step1: Create Control Set and Step 2: Create Additional Sample.
2. If your expert reviewers have approved the control set, then check **The Control Set has been reviewed** option.

3. Click **Run Test** to apply predictions to all of the items in the control set.



Once you click "Run Test", a job starts which you can monitor from the Jobs window. The system generates a report with prediction results. When ready, you can click **View Report** to view the results.

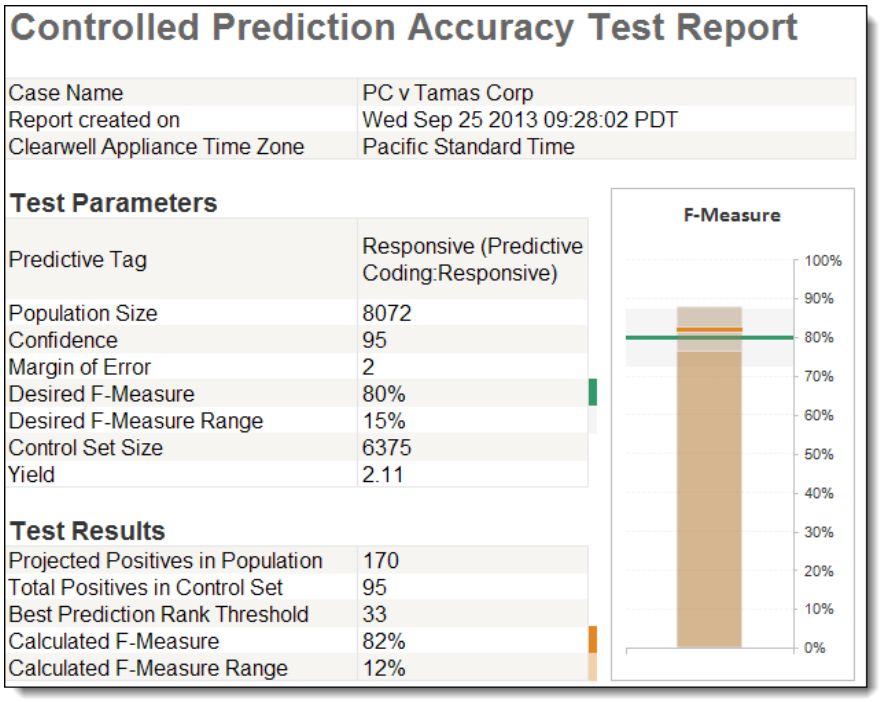
4. Click **View Report** to open and view the prediction results.

Controlled Prediction Accuracy Test Report

The Controlled Prediction Accuracy Test report is a great way to get insight into your data and quickly identify where you are in comparing human tagging decisions against the system's prediction ranks.

Controlled Prediction Accuracy Test Report Section

SUMMARY



Controlled Prediction Accuracy Test Report

Case Name	PC v Tamas Corp
Report created on	Wed Sep 25 2013 09:28:02 PDT
Clearwell Appliance Time Zone	Pacific Standard Time

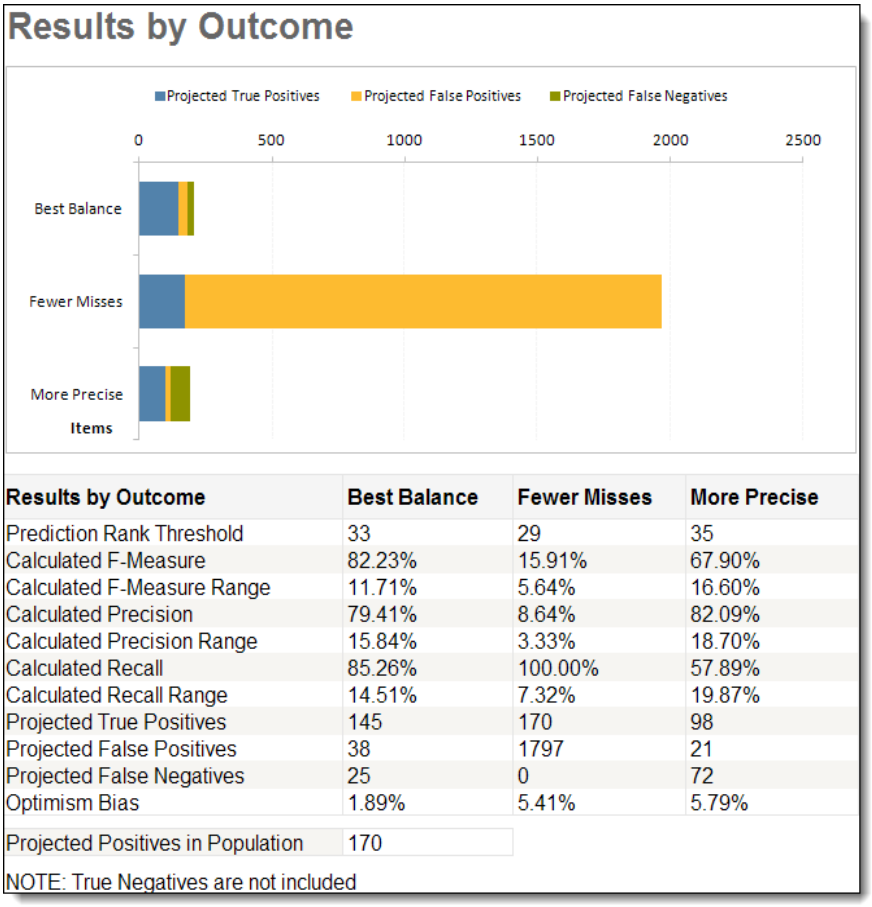
Test Parameters

Predictive Tag	Responsive (Predictive Coding:Responsive)
Population Size	8072
Confidence	95
Margin of Error	2
Desired F-Measure	80%
Desired F-Measure Range	15%
Control Set Size	6375
Yield	2.11

Test Results

Projected Positives in Population	170
Total Positives in Control Set	95
Best Prediction Rank Threshold	33
Calculated F-Measure	82%
Calculated F-Measure Range	12%

Results by Outcome



Results by Outcome	Best Balance	Fewer Misses	More Precise
Prediction Rank Threshold	33	29	35
Calculated F-Measure	82.23%	15.91%	67.90%
Calculated F-Measure Range	11.71%	5.64%	16.60%
Calculated Precision	79.41%	8.64%	82.09%
Calculated Precision Range	15.84%	3.33%	18.70%
Calculated Recall	85.26%	100.00%	57.89%
Calculated Recall Range	14.51%	7.32%	19.87%
Projected True Positives	145	170	98
Projected False Positives	38	1797	21
Projected False Negatives	25	0	72
Optimism Bias	1.89%	5.41%	5.79%
Projected Positives in Population	170		

NOTE: True Negatives are not included

The report, which is divided up into summary and Results by Outcome, lists the test parameters, test results, and displays the F-Measure and Results By Outcome data in table and bar chart format.

Note: The same reporting format is used for the Prediction Accuracy Test report. See *"Prediction Accuracy Test Report" on page 53*. For report details, see *"How does the Controlled Prediction Accuracy Test differ from the Prediction Accuracy Test?" on page 77*.

Table 7: Controlled Prediction Test

Field	Description
Case Name	Name of the case
Report created on	Date and time of report
Appliance Time Zone	Time Zone of the appliance
Test Parameters	
Predictive Tag	Name of predictive tag
Population Size	The total number of items.
Confidence	Input value (percentage) for Step 1.
Margin of Error	Input value (percentage) for Step 1.
Desired F-Measure	Goal or desired value (percentage) which corresponds to **green** bar in F-Measure graph. For example, 80%.
Desired F-Measure Range	Desired range (percentage) which corresponds to the shaded region in the F-Measure graph. For example, 10% would show 5% above and below the Desired F-Measure.
Control Set Size	The actual item count in both sample folders.
Yield	Calculated value for Step 2.
Test Results	
Projected Positives in Population	The number of projected predictive tag positives for the entire population
Total Positives in Control Set	The number of items that were tagged with your **Select Tag** option. **Note:** This number only relates to the control set.
Best Prediction Rank Threshold	An index into the report details section. For example, if the value is set to 70, locate that value under the Threshold column in the Report Details (next section).
Calculated F-Measure	Value (percentage) currently achieved which corresponds to the **orange** line in the F-Measure graph.
Calculated F-Measure Range	Range value (percentage). On the F-Measure graph, the area is shaded to highlight the Calculated F-Measure range. For example, 7% would show 3.5% above and below Calculated F-Measure.

Results By Outcome Section

Table 8: Results By Outcome

Different Threshold Options:	
Best Balance	Prediction rank threshold with the highest possible calculated F-Measure (or balance between precision and recall)
Fewer Misses	Prediction rank threshold with 10% higher recall than Best Balance (or a higher recall than the Best Balance, but at a loss of precision)
More Precise	Prediction rank threshold with 10% higher precision (lower recall) than Best Balance (or a higher precision than the Best Balance, but at a loss of recall)
Results By Outcome	
Threshold	Prediction rank threshold
Calculated F-Measure	Harmonic mean of calculated precision and recall.
Calculated F-Measure Range	Calculated F-Measure range percentage.
Calculated Precision	Percentage of all items above threshold that is positive
Calculated Precision Range	Calculated precision range
Calculated Recall	Percentage of all positives in sample above the threshold
Calculated Recall Range	Calculated recall range
Projected True Positives	Estimated number of items above the prediction rank threshold that would be tagged with the predictive tag.
Projected False Positives	Estimated number of items above the prediction rank threshold that would not be tagged with the predictive tag.
Projected False Negatives	Estimated number of items below the prediction rank threshold that would be tagged with the predictive tag.
Optimism Bias	Potential difference between expected results and actual outcome as a result of repeated iterations. For example, if the Calculated F-Measure was 80% with an Optimism Bias of 1%, then there is a statistical chance that the Calculated F-Measure would have been 79% if tested against a newly selected random sample. **Note:** The optimism bias row is shown in the report only if its value is 10% more than the corresponding F-Measure range.
Projected Positives in Population	Estimated number of items in the population that would be tagged with the predictive tag

Training the System

The system uses machine learning to predict tag decisions for individual tags. Selecting the action Train teaches the system to recognize patterns in your reviewer's tagging decisions.

This example uses a subset of training items from the Training Set 1 folder and the Responsive tag to build the Transparent Predictive Coding system.

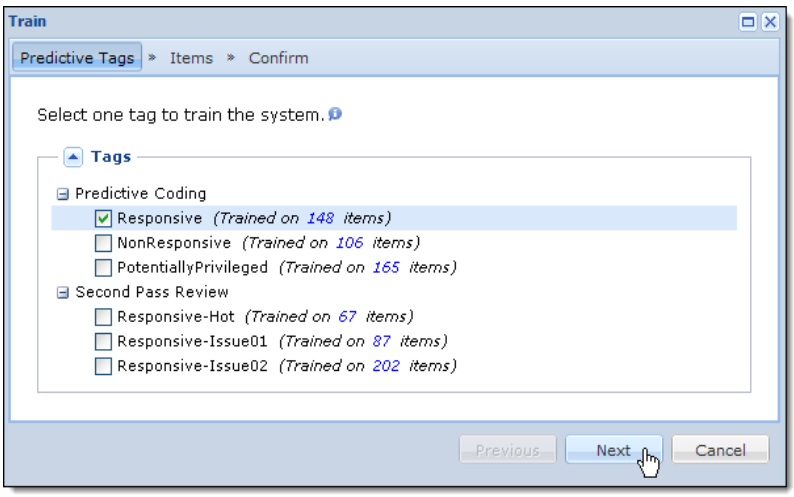
	Important:	Training clears all the previous prediction ranks.
	Restrictions:	Maximum number of tags you can train on is 20. Maximum number of items you can train on is 10,000.
	Inputs:	The reviewed and tagged training set items. **Note:** Training items must be viewable within current search.
	User:	Case Administrator and Case Managers

Step 1: Select Predictive Tag

The first step of the Train process is to select a tag to train the system. The system learns to recognize patterns in the items for a given tag and then predicts how likely other documents are to be tagged. For example, if you train on Responsive, you can predict how likely the other items will also be Responsive.

To select a Predictive Tag

1. From the Documents Screen under the Analysis & Review module, click **Action** and select **Train** to start training the system.
2. Select a tag for the system to learn.



Note: If this is not the first training iteration, then the previously trained item count is shown.

3. When you have finished selecting the tag, click **Next**.

Step 2: Select Items

Once you have selected the predict tag, the **Select Items** screen displays, requesting items to train on.

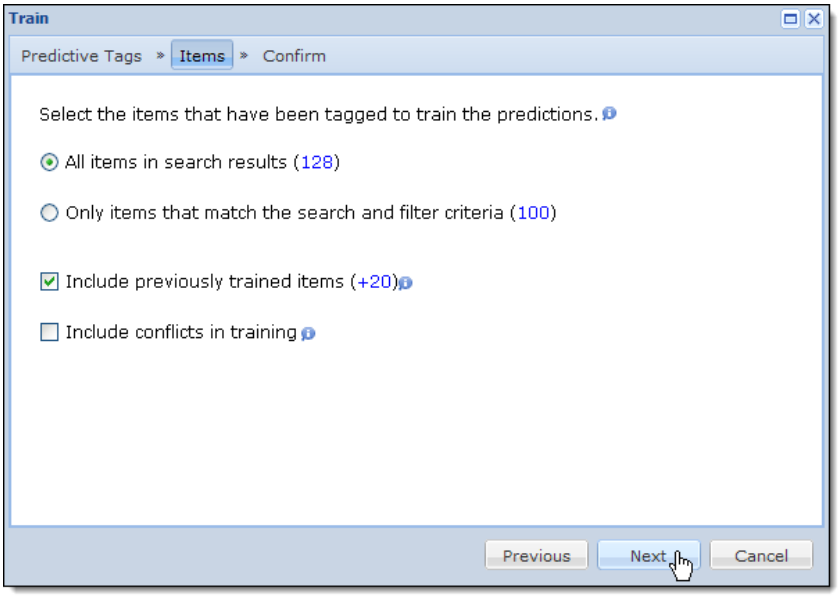
The system learns on an individual item basis. When you specify the items to train on, it is important to select all items that have been reviewed and not just the ones that have the tag that is being learned. For example, if you are training on the Responsive tag, you also want to include everything that has been reviewed and is Not Responsive.

To select Items

1. Specify the following information, depending on the type of training:

Table 9: Train - Step 2: Select Items

Field	Description
All the items in the search	Includes every email message, loose file, attachment, and embedding even if it did not match the search.
Only items that match the search and filter criteria	Only includes the items which match the search criteria and are tagged. An empty corpus search will cause every item to match.
Include previously trained items	Training involves building a learning model from all of the training items provided. It is not incremental, it re-learns each time. In most situations select this option unless you want to start over.
Include conflicts in training	Tagging decisions that mark duplicate items differently are identified by the system as conflicts. For example, if "attach.doc" is tagged Responsive, and another email message also has "attach.doc" but it was tagged Not Responsive, then the items are found to be in conflict. Including items that are marked as conflicts without resolving the underlying reason for the conflict can affect and potentially lower the prediction accuracy. **Important:** Veritas recommends resolving conflicts and then re-training the system to increase prediction accuracy If conflicts are not included, the system ignores all conflicts. A conflict report is available for download once the training job completes. See *"View Job and Training Error Report" on page 42*.



2. After you have finished selecting items, click **Next**.

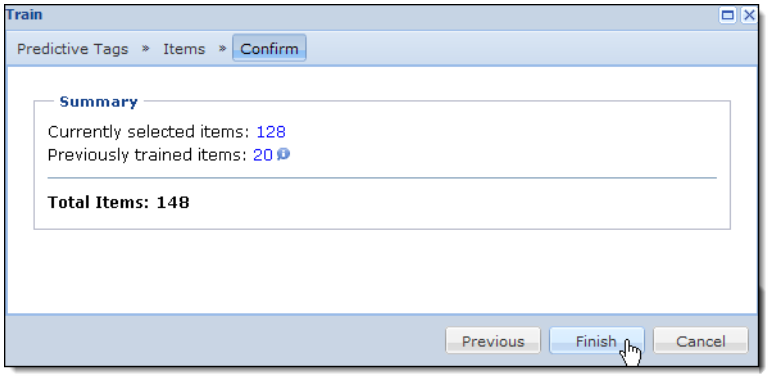
Step 3: Confirm Training Selections

The confirmation window shows a summary and total of the current selected items and previously trained items. If you determine that this training set performs well, you can move on to the next training set.

Note: An error displays if the system does not have enough items or has too many similar items selected. Go back and correct the composition of items in order to continue and submit the training job.

To confirm training items

- Confirm the total items to train on and click **Finish** to submit the training job.

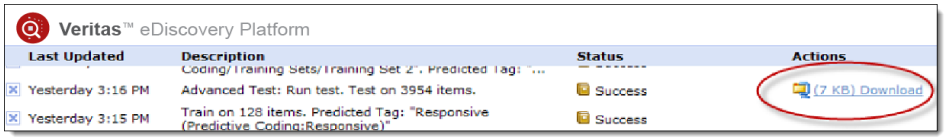


View Job and Training Error Report

A conflict is any contradictory set of data or tagging decisions. The conflict report is generated as part of the training job. It describes the conflict in terms of the training set and the Transparent Predictive Coding system.

To view the Train results and Error Report

- The Jobs window displays the status of the train job displays. Under Actions, click the link to view details for the errors or conflicts that occurred during this phase.



- The following sample Train Error Report provides a description of possible errors.

  Note: A feature is a property of a document or item representing content or metadata.

Conflict Type	Resolution	Document IDs			
Similar Features	DROP	0.7.11.7236	0.7.11.8314-000001	0.7.11.13198	
Similar Features	DROP	0.7.11.14885	0.7.11.11731	0.7.11.9502-000001	
No Features	DROP	0.7.11.18447	0.7.11.15627	0.7.11.16634	0.7.11.18445
Similar Features	DROP	0.7.11.12982	0.7.11.11451		
Similar Features	DROP	0.7.11.14476-000001	0.7.11.10956		
Similar Features	DROP	0.7.11.16601	0.7.11.11117		

Table 10: Understanding the Training Error Report

Error Type	Description	Recommendation
Similar Features	If a group of items with similar features tagged inconsistently then it is considered as conflicting tag decision.	The entire group of items is dropped if there is a conflict. If you specified the ignore conflicts option in the Train action, you can ignore this error.
No Features	If the system could not extract any features from the item.	These items are always dropped from training. Please resolve these errors before continuing on. Some tips to consider: • If this error is due to TIFF or PDF images, then please process the images using OCR to extract content from the items. • There may be some cases (such as an executable files) where there is no resolution. These items will be excluded from the predictive coding process.
Excluded items	Certain type of items like "EXE" and "DLL" are excluded.	These items are always dropped from training.

Predicting

The Predict action computes a ranking value or score that equates to how likely an item is to be tagged with a selected tag. This computation examines the meta-data and content of prediction items and determines the set of training items with similar meta-data and content that are present in the system. The ranking value range is from 1 to 100 with 1 being least likely and 100 being most likely to be tagged.

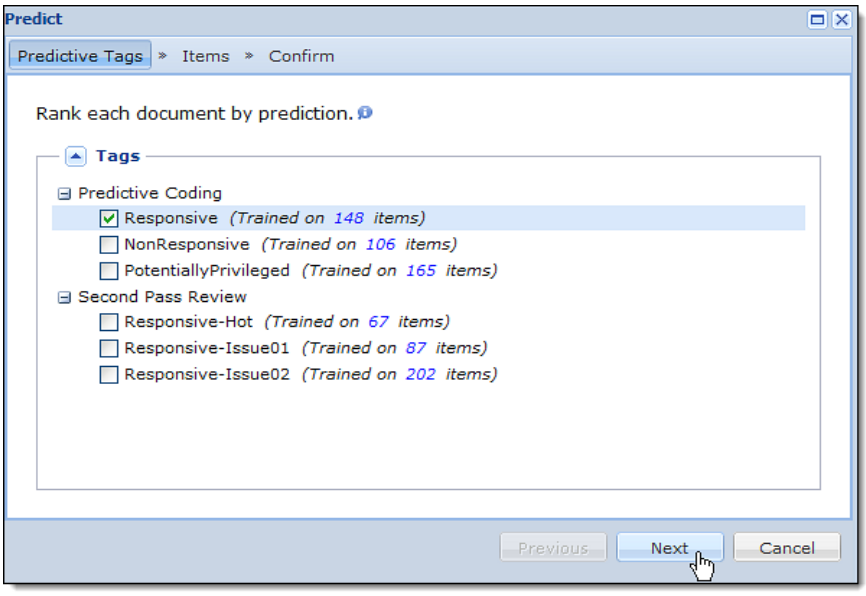
Note: When the predict job completes, re-run your search to see additional features, including advanced search options, filters, and ranks for each document.

	Important:	Do not apply Predict to items that were used for training.
	Inputs:	Any set of items viewable in current search. Predictive tags
	User:	Case Administrator and Case Managers

Step 1: Select Predictive Tag

To select predictive tag

1. Select a tag from the list.



2. Click **Next** to advance to Select Items.

Step 2: Select Items

In order to provide flexibility on the set of items to be predicted, the product does not automatically apply ranks to every item. Instead, you specify on the Select Items screen the items for prediction. You should select items for which you want to see prediction ranks but have not been previously selected to train the system (such as the next test set, next review set, or the entire item population).

To select the items that will be predicted

1. Specify the items for prediction.

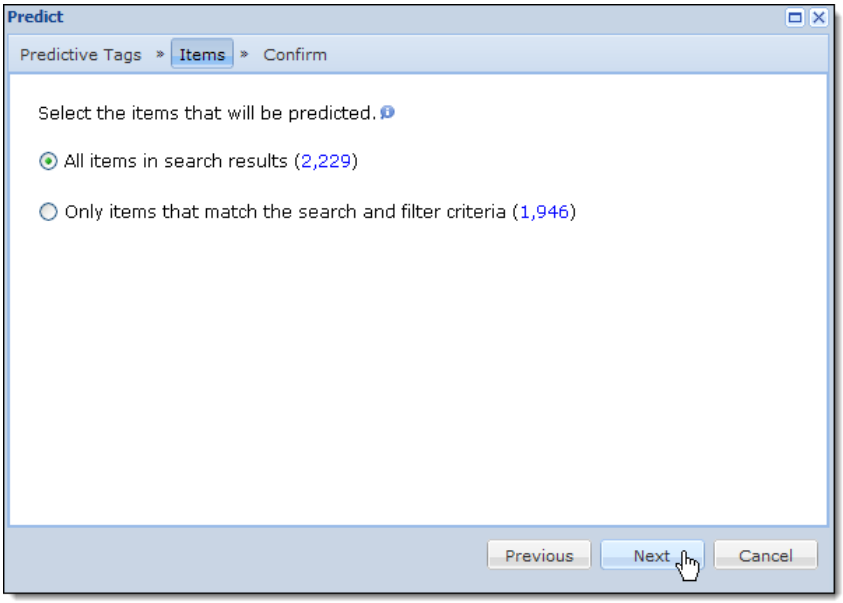


Table 11: Predictive Tag - Step 2: Select Items

Field	Description
All items in search results	All items in the search results includes every email message, loose file, attachment, and embedding even if it did not match the search. (Recommended)
Only items that match the search and filter criteria	Only items that match the search and filter criteria will only include the items which match the search criteria. An empty corpus search will cause every item to match

2. Click **Next** to confirm selections.

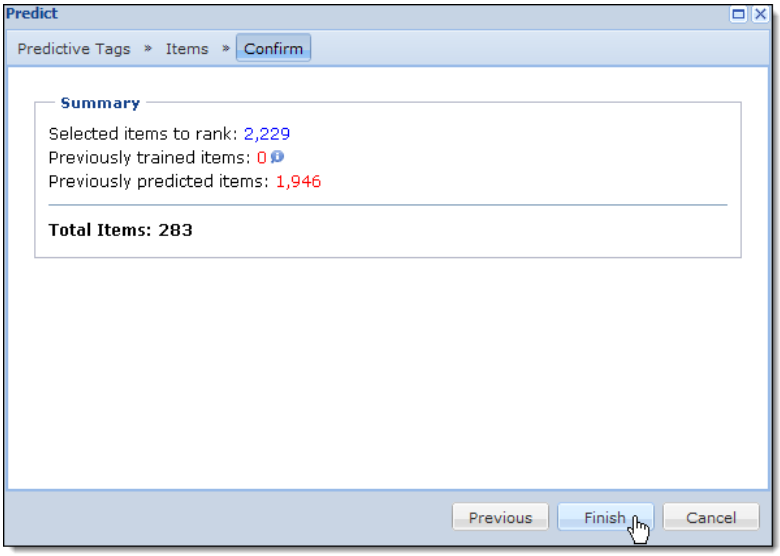


Table 12: Predictive Tag - Item Summary

Field	Description
Selected items to rank	Includes items in the current search results.
Previously trained items	Prediction ranks are not applied to documents which have been used for training. These items are not included in the Predict action.
Previously predictive items	The total number of items to be predicted.

3. Click **Finish** once you validate the item summary.

Selecting the Next Training Set

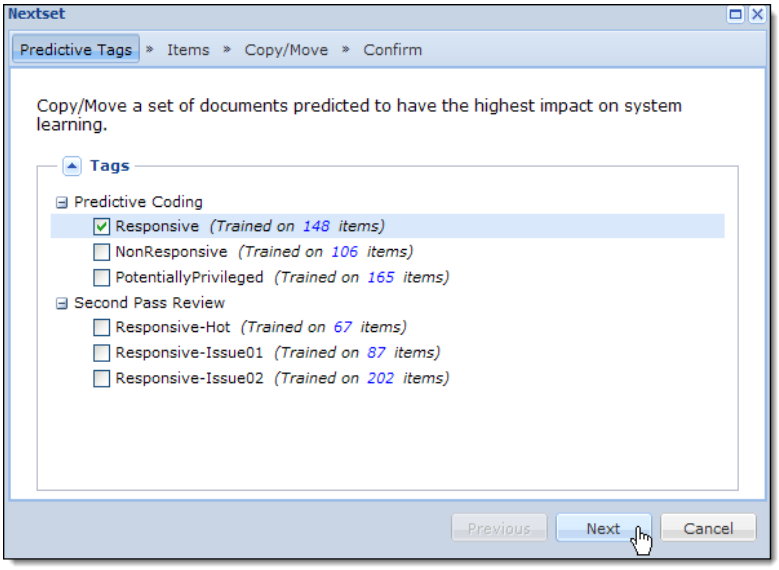
If the Transparent Predictive Coding system has not learned enough to predict and meet desired recall and precision metrics, the action Next Training Set allows you to generate the next set of items for the system to learn and improve its prediction capabilities. This action allows you to copy or move items within the Transparent Predictive Coding workflow.

	Important:	Typically, items that comprise the Next Training Set have *not* been reviewed.
	Inputs:	Predictive tags Population either viewable in current search or in a predictive coding specific folder.
	User:	Case Administrator and Case Managers

Step 1: Select a Next Training Set Tag

To select Next Training Set tag

1. From the Documents Screen under the Analysis & Review module, click **Action** and select **Next Training Set**. The Next Training Set screen displays.
2. Select the tag for which you want to improve accuracy.



3. When you have finished selecting the tag, click **Next** to advance to the Items screen.

Step 2: Select Items for Next Training Set

1. Select a batch of items from the population to comprise the next training subset.

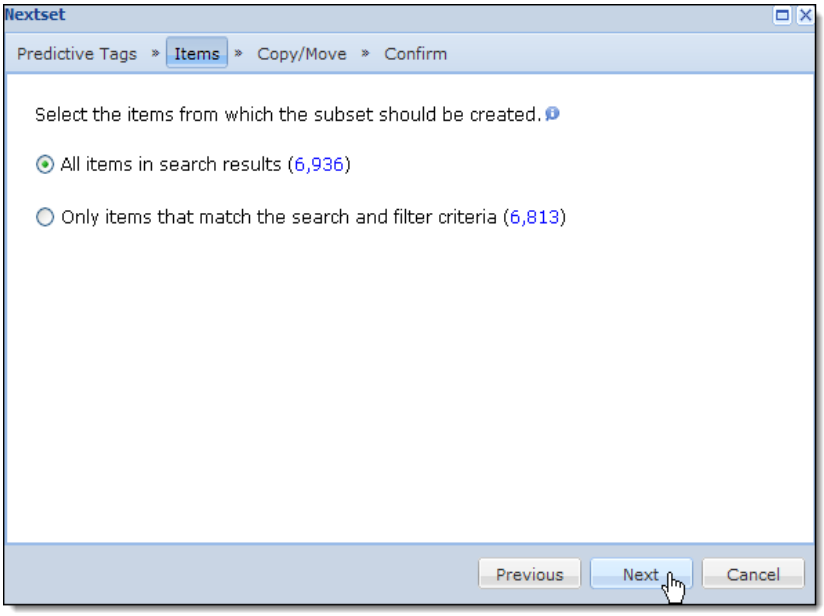


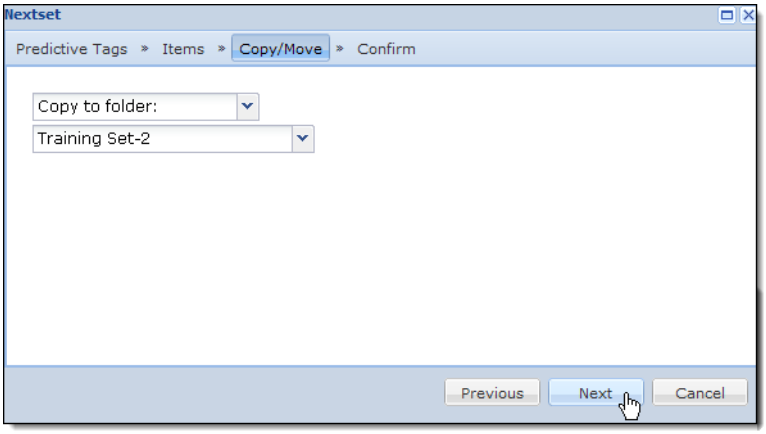
Table 13: Next Training Set - Step 2: Set Items

Field	Description
All items in search results	Includes every email message, loose file, attachment, and embedding even if it did not match the search.
Only items that match the search and filter criteria	Include the items which match the search criteria. An empty corpus search will cause every item to match.

2. Click **Next** to advance to the Copy/Move screen.

Step 3: Copy or Move Items to Folder

1. Click the first drop-down menu to select whether you want to copy or move items to a folder.



2. Click the second drop-down menu to select the appropriate training folder and click **Next**.

Step 4: Confirm the Next Training Set Items

1. Verify the information and when ready, click Confirm to launch the Next Set job.

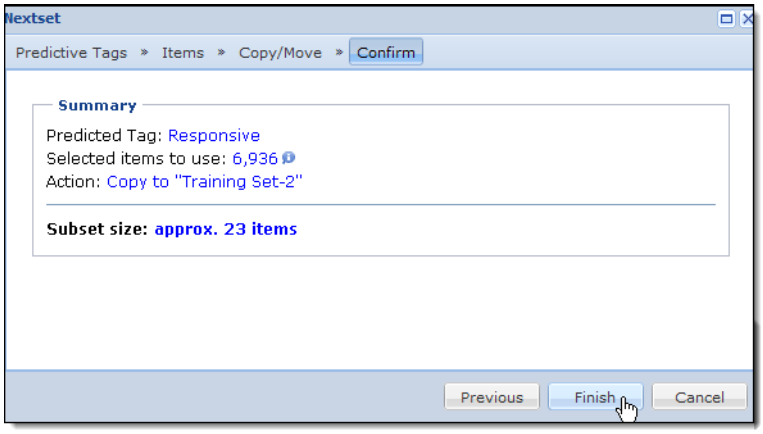


Table 14: Next Training Set - Step 4: Confirm Items

Field	Description
Predictive Tag	Displays the selected predicted tag.
Selected items to use	These items are used for the Next Set.

Table 14: Next Training Set - Step 4: Confirm Items

Field	Description
Action	Either: • Copy to folder • Move to folder
Subset size	The approximate size of the next training review set.

Prediction Accuracy Test

Prediction Test allows you to quickly test the current learning state of the system against a random set of items that have been reviewed. The role of Prediction test differs from Controlled Prediction Accuracy Test in that it is not as comprehensive and does not utilize or measure against the controlled and statistically robust Controlled Prediction Accuracy Test. For full defensibility and transparency of your work product, you should use Controlled Prediction Accuracy Test.

Prediction Test is flexible and can be very useful if you have devised your own sampling technique. It is not a replacement for the Controlled Prediction Accuracy Test but rather a complimentary testing tool.

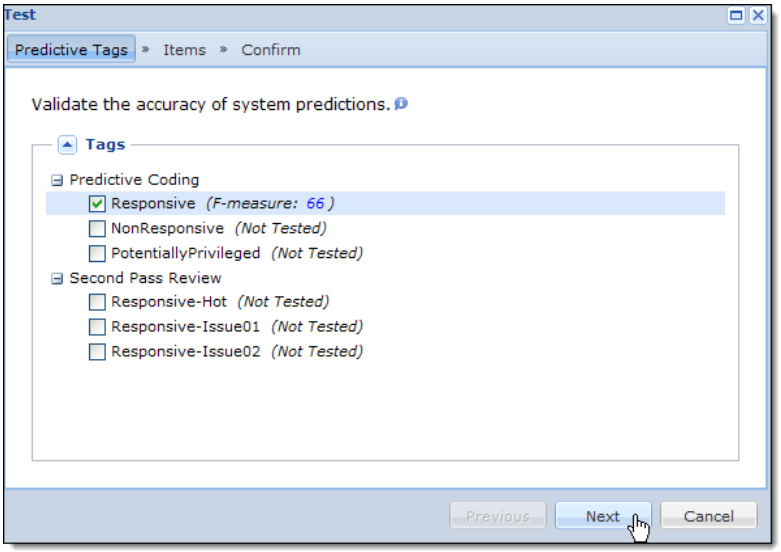
	Important:	Prior to starting the Prediction Accuracy Test, all items must already have Prediction ranks (Predict action).
	Inputs:	Items viewable in current search results.
	User:	Case Administrator and Case Managers

Step 1: Select Prediction Test Tag

The system will compare the human tagging decisions against the system's own prediction ranks and provide a report. For each rank, the report will tell you how many items were predicted correctly and what you can expect if you were to bulk tag the remaining items.

To select Prediction Test tag

1. From the Documents Screen under the Analysis & Review module, click **Action** and select **Prediction Test**. The Prediction Test screen displays.



2. Select a tag to validate the accuracy of system predictions.

3. Click **Next** to advance to the Items screen.

Step 2: Select Items

Prediction Test relies on and works best with a statistically valid sample. Item selections should be completely random, so you may want to select only the items that match your search and filter criteria (depending on how you set up your test sample).

1. Select the Prediction Test items

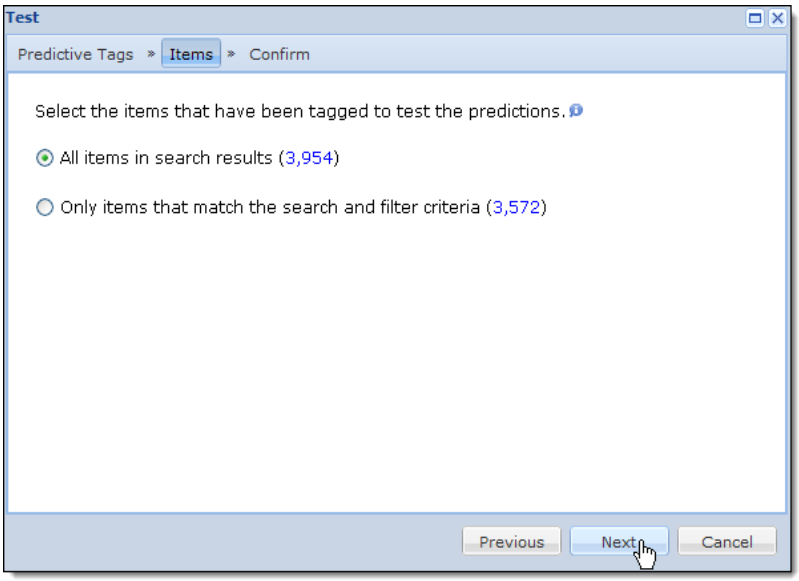


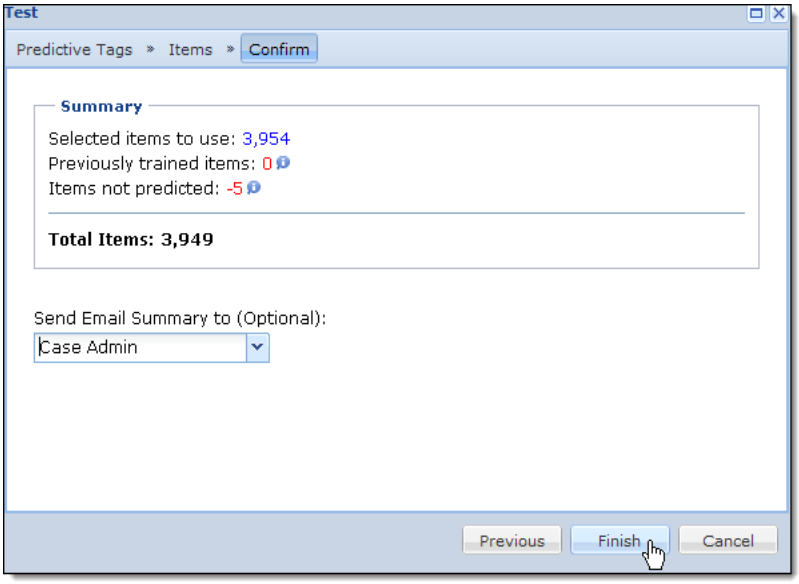
Table 15: Prediction Test - Step 2: Select Items

Field	Description
All items in the search results	Includes every email message, loose file, attachment, and embedding even if it did not match the search.
Only items that match the search and filter criteria	Only include the items which match the search criteria. An empty corpus search will cause every item to match.

2. Click **Next** to advance to confirmation menu.

Step 3: Confirm Prediction Test Item Selections

1. View the summary of selections. You also have the option of notifying your review team of the summary.



   Note: If there were any issues or errors preventing the Prediction Test, an explanatory warning message displays at the top of the screen.

2. Optional: You can get an email notification when the job completes. From the **Send Email Summary to (Optional)** drop-down menu, select the email recipient.

3. If the summary list is complete, click **Finish**. A Prediction Test job starts and a report is available for pickup in the Jobs window. You can also view the report from the Prediction Status page.

Step 4: View the Prediction Test Report

Prediction Accuracy Test Report

Similar to the Controlled Prediction Accuracy Test report, the Prediction Accuracy Test report keeps track of the prediction ranks, making it easy for you to learn precisely what is happening and the progress you are making toward your sampling goal.

Note: Since the Prediction Accuracy test does not does not follow the same rigorous training cycle sequence as the Controlled Prediction Accuracy Test, there are fewer reporting metrics. The minimum and maximum of Recall, Precision and F-Measure metrics are not calculated.

Example of Report Output

Prediction Accuracy Test Report & Results By Outcome

A report page provides Prediction Accuracy test results, test parameters and a table and bar chart of the Results By Outcome. You can drill deeper into the data by viewing the various worksheets (Summary, Results by Outcome and Accuracy Metrics) of the report. For a description and explanation of the various fields that are displayed in the report, see *"Controlled Prediction Accuracy Test Report Section" on page 35*.

Prediction Accuracy Test Report

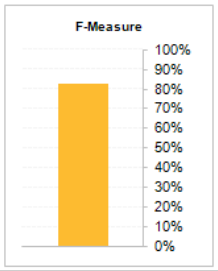
Case Name	PC v Tamas Corp
Report created on	Wed Sep 25 2013 09:33:29 PDT
Clearwell Appliance Time Zone	Pacific Standard Time

Test Parameters

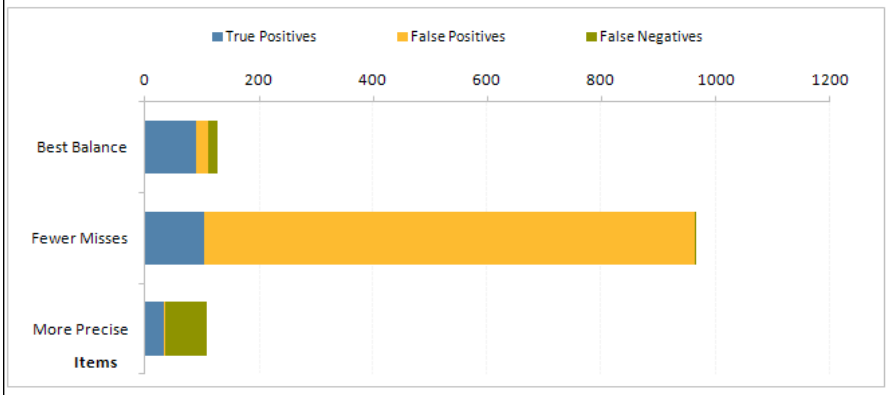
Predictive Tag	Responsive (Predictive Coding:Responsive)
Test Set Size	4541

Test Results

Total Positives in Test Set	105
Best Prediction Rank Threshold	33
Calculated F-Measure	82%



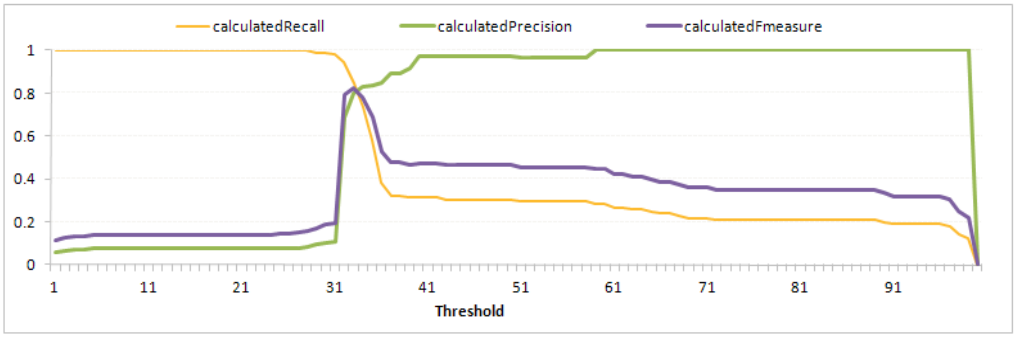
Results by Outcome



Results by Outcome	Best Balance	Fewer Misses	More Precise
Prediction Rank Threshold	33	31	39
Calculated F-Measure	82.41%	19.27%	46.81%
Calculated Precision	80.18%	10.68%	91.67%
Calculated Recall	84.76%	98.10%	31.43%
True Positives	89	103	33
False Positives	22	861	3
False Negatives	16	2	72

NOTE: True Negatives are not included

Accuracy Metrics for Prediction Rank Threshold



thresh	numD	numTr	numFal	numFal	calculatedRec	calculatedPrec	calculatedFme	Outcome
20	1389	105	1284	0	1	0.075593956	0.140562251	
21	1387	105	1282	0	1	0.075702958	0.140750676	
22	1386	105	1281	0	1	0.075757578	0.140845075	
23	1386	105	1281	0	1	0.075757578	0.140845075	
24	1375	105	1270	0	1	0.076363638	0.141891882	
25	1364	105	1259	0	1	0.076979473	0.142954379	
26	1346	105	1241	0	1	0.078008913	0.144727767	
27	1311	105	1206	0	1	0.080091536	0.148305103	
28	1250	105	1145	0	1	0.083999999	0.154981554	
29	1114	104	1010	1	0.990476191	0.093357272	0.170631677	
30	1018	104	914	1	0.990476191	0.102161102	0.18521817	
31	964	103	861	2	0.980952382	0.106846474	0.192703471	Fewer Misses
32	144	99	45	6	0.942857146	0.6875	0.795180738	
33	111	89	22	16	0.847619057	0.801801801	0.82407409	Best Balance
34	93	77	16	28	0.733333349	0.827956975	0.777777791	
35	73	61	12	44	0.580952406	0.83561641	0.685393274	
36	47	40	7	65	0.380952388	0.851063848	0.526315808	
37	38	34	4	71	0.323809534	0.894736826	0.475524485	
38	38	34	4	71	0.323809534	0.894736826	0.475524485	
39	36	33	3	72	0.314285725	0.916666687	0.46808514	More Precise
40	34	33	1	72	0.314285725	0.970588207	0.474820167	
41	34	33	1	72	0.314285725	0.970588207	0.474820167	

Managing Transparent Predictive Coding

This section describes the various management tasks that you can perform to track and monitor the progress and results of Transparent Predictive Coding workflows.

Refer to the following topics in this section:

Case Settings for Transparent Predictive Coding

Starting with 7.1.2, Transparent Predictive Coding is automatically enabled. It is highly recommended that you do not disable the feature even if you are not using it.

Note: If you had previously disabled Transparent Predictive Coding for a case, and you want to enable it, make sure that you enable the feature and rerun post-processing.

For more information on how to enable or disable the feature, refer to the Case Administrator Guide "Enable Predictive Coding".

Predictive Tag and Training Items Management

There are two types of limits in the Train action. First, the number of predictive tags cannot exceed 20 and secondly, the maximum number of items cannot exceed 10,000. If you have more than 20 predictive tags, you can use the Export Training function for a predictive tag that is not needed and then use Delete Training to allow you to train on another predictive tag.

Prediction Status

The Prediction Status page provides a listing of all the actions and activity performed on predictive tags. You can view current training cycle information and see how various actions and events ran in the Transparent Predictive Coding system and who ran them. There are two types of information displayed:

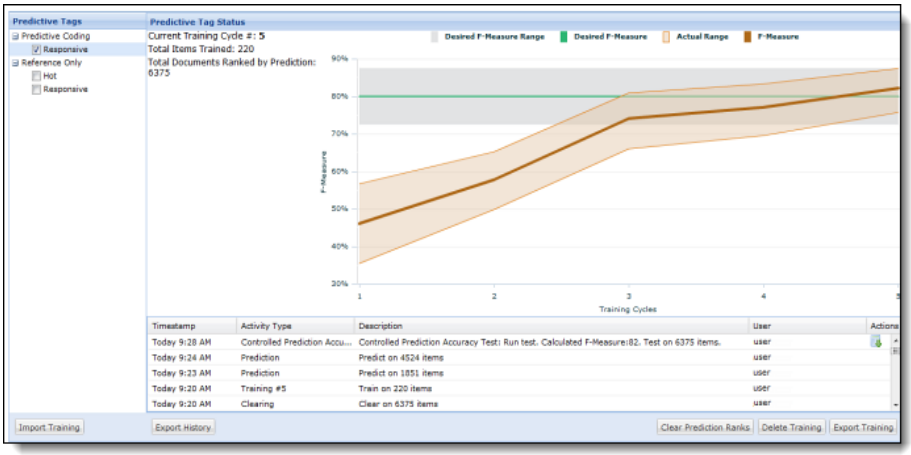
- **Predictive Tags** — Lists all the available predictive tags
- **Predictive Tag Status** —Record of events on the selected predictive tag. These events include a timestamp, activity type, description, user and a report.

Viewing Predictive Tag Information

To view predictive tags and predictive tag status

1. From the **Analysis & Review** module, select **Predictive Status** to view predictive tag activity.

   The Predictive Status screen shows the status (right column) and lists the predictive tag (left column). A helpful visualization tool, the Predictive Tag Status chart, tracks the progress of the Predictive Tag in relation to the number of training cycles.



2. On the Predictive Tags screen, all of the predictive tags are displayed. For import training information, see *"Import Training" on page 60*.

3. On the Predictive Status chart, 4 dimensions are plotted. Over time and after multiple runs, you should notice the progress of the training cycles on your data. Hovering over each data point, provides the Training Cycle, F-Measure, Actual Range and either the size of the Control Set for the Controlled Prediction Accuracy test or the size of the Test Set for the Prediction Accuracy test for each data point.

   Note: The Prediction status screen displays the chart only after completing two training cycles.

Table 17: Predictive Status Chart

Field	Description
Desired F-Measure Range	Target range of acceptable F-Measure values (gray). This is generated from the input from Step 2 of the Controlled Prediction Test (gray).
Desired F-Measure	Measures the accuracy of a test sample (green))
Actual Range	Calculated F-Measure range for iteration (beige)
F-Measure	Calculated F-Measure value for iteration (brown)

4. The Predictive Tag Status column displays the following record information for each entry:

Table 18: Predictive Tag Status

Field	Description
Timestamp	Date and time when the action completed.
Activity Type	The following activity types are recorded here: • Controlled Prediction Accuracy Test • Train • Predict • Prediction Test • Next Training Set • Clear • Import • Export
Description	• Text summary of the activity.
User	Displays the user who ran the action.
Actions	Downloadable reports for the following activity types: • Controlled Prediction Accuracy Test: Test Report • Train: Conflict Report (if applicable) • Prediction Test: Test Report Run Search (click magnifying glass icon) for only: • Initial sample items • Additional sample items **Note:** If you clear the control set and build it again, there will more than on entry for the initial and additional sample items for the associated tag in the prediction status page. In such cases, the magnifying icon will appear only for the latest revisions of the initial and additional sample items. The previous versions of the initial and additional samples are not searchable.

5. You can perform the following actions from the Predictive Tag Status screen.

Table 19: Predictive Tag Status Actions

Field	Description
Export History	Export the history report to a CSV file.
Clear Predictions	Clear predictions for a given tag.
Delete Training	Delete training for a predictive tag and clears all history.
Import Training	Import predictive tag data to the system.
Export Training	Export predictive tag data to the system.

Import Training

Import Training allows you to leverage training from other cases. For example, if privilege issues are similar across cases you could immediately set aside privileged items. Similarly, you may want to import training in order to utilize Transparent Predictive Coding for more effective culling of spam email.

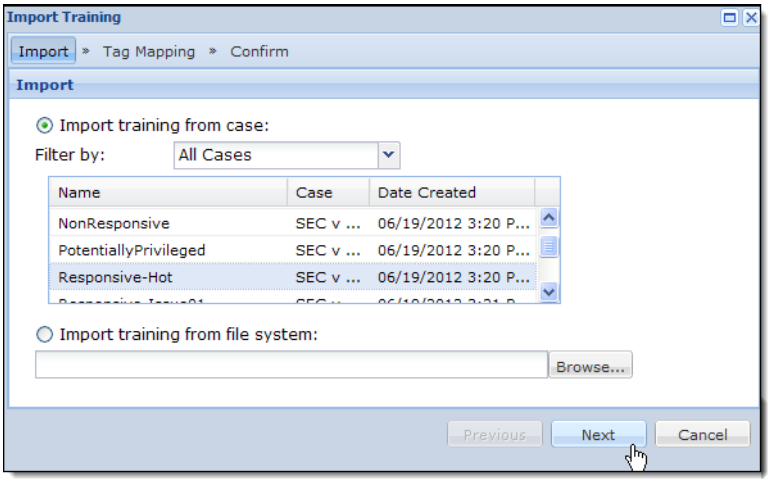
During this phase, only the Predictive Tag is imported, and not the documents that were used for training to create it. This means that subsequent training only uses new documents from the current case and you will not see any information displayed in the prediction insight or exemplars since the training documents are not available.

Important: When importing training, only tags from cases that you have access to are shown. Similarly, the system only displays tags that were exported from the original case.

To import training

1. From the Predictive Tags screen, select **Import Training**.

   Import Training screen opens.

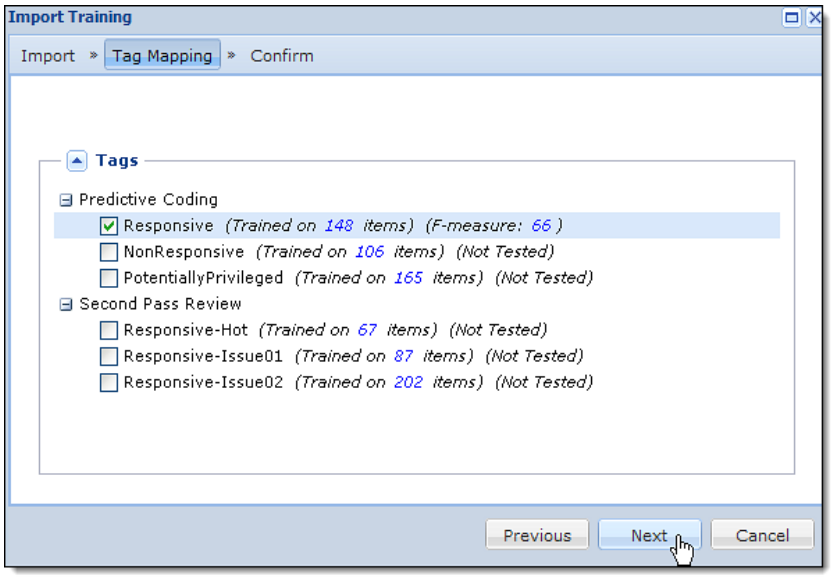


   Note: Import training from file system refers to the operating system's file system and not to an upload feature of the Transparent Predictive Coding platform. Typically, these are cases from outside the current cluster.

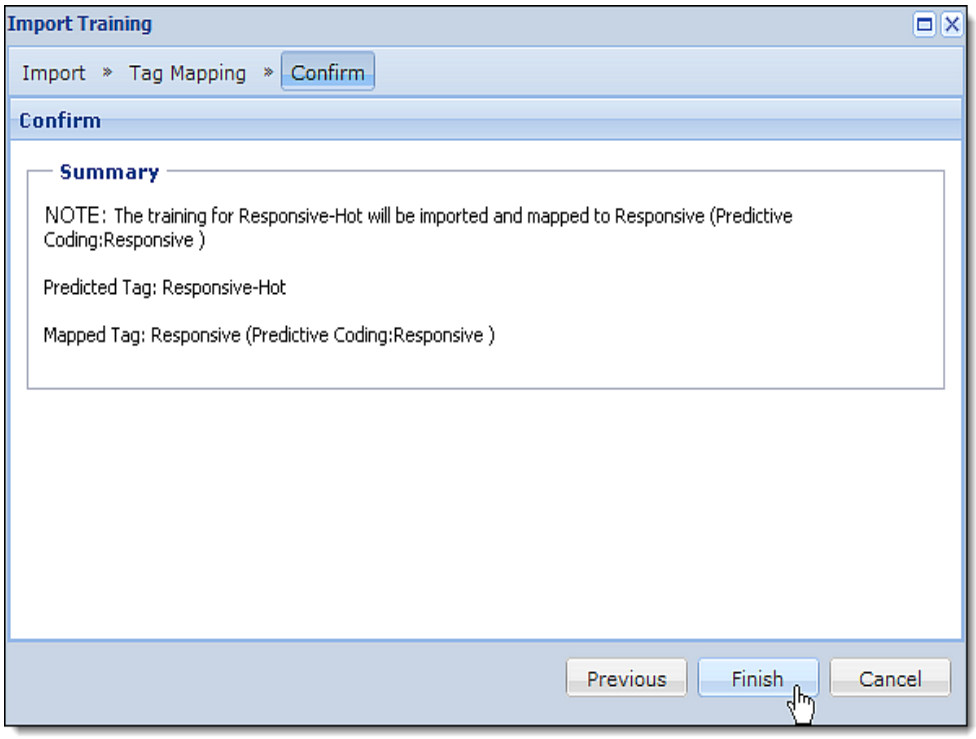
2. Select whether you want to import your training from a case using the drop-down menu or file system and click **Next**.

3. Map the imported training to a tag and click **Next** to advance to the Confirm screen.

   For this example, the Responsive tag is selected.



4. On the Confirm screen, verify the selections and click **Finish** to start the Import Training job.



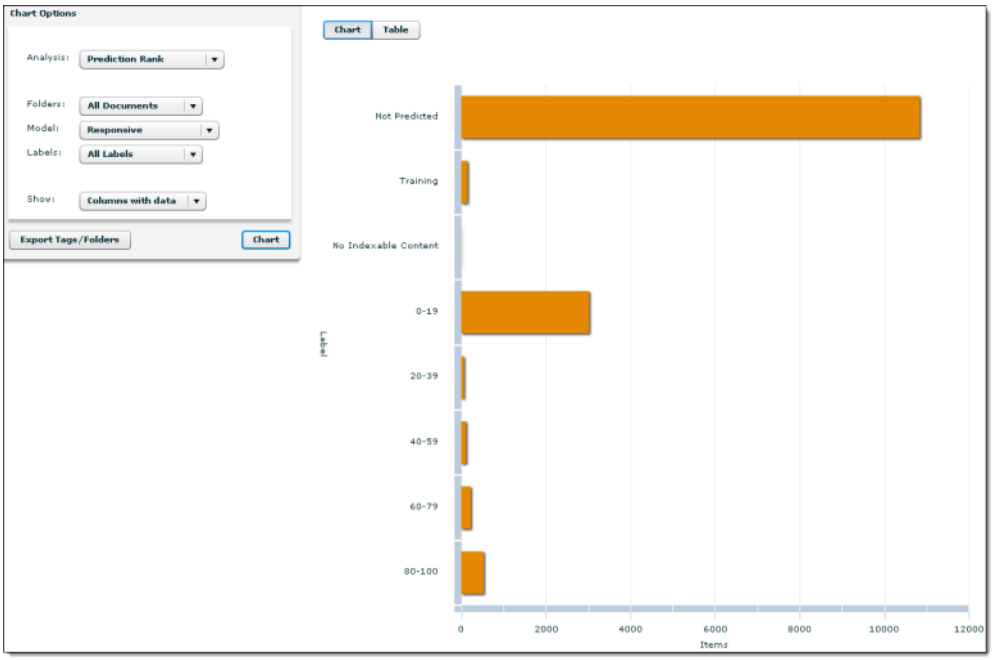
Note: Importing training imports the learning from the training and not the physical items from training data set. This means that you can execute the Predict and Next Training Set actions but not the Train action.

Using the Review Dashboard

The Review Dashboard provides a set of interactive charts and reports that measure prediction accuracy and analyze items by probability score. Case Administrators and reviewers (with appropriate permissions) can view Prediction Rank information in either chart or table format.

To access Prediction Rank information from the Dashboard

1. From **Analysis & Review**, in the Review Dashboard Chart Options box, select Analysis: Prediction Rank.



2. From the Folders drop-down list, select all documents, all folders, or a specific folder to display data.
3. Select a predictive tag to view from the Models menu. This example shows the predictive tag Responsive.
4. From the Labels drop-down list, select all labels or a specific label to display filter criteria data.
5. Select whether you want unused columns to display in the chart.
6. Click **Chart**.

   A chart displays showing the number of prediction tag items by filter criteria.
7. To view the displayed information in table format, select the **Table** view.

Granting/Restricting Transparent Predictive Coding Access

As a Case Administrator or Manager you can allow or restrict the viewing of prediction ranks, actions and management views. For information on granting or restricting Transparent Predictive Coding access, refer to *"Administering User Accounts" in the System Administration Guide*.

Predictive Analytics with Search and Review

To better understand and interpret the Transparent Predictive Coding results, the Search and Review capabilities display prediction ranking information. You are already familiar with the benefits of Advanced Search and Review on the platform and the prediction ranking additions are logical extensions to assist in review analysis and case strategy. These analytics make sense of the available prediction results using a bar ranking, action status and percentage prediction metric and are designed to inform your next step in the review workflow process.

Refer to the topics in this section:

Viewing Prediction Rank Search Results

Starting with 7.1.2, a *Prediction* column has been added to the search results page. This new column allows you to select a single tag, sort, and view a bar chart or a status icon of the prediction ranking for all items that have been included in the building of the system. You can examine prediction ranks at the parent item level or on any single item.

Note: Sorting the column uses the "family score" and not the score for the email item itself. This means that when you sort, the bars will not necessarily (and in fact will probably not) appear to be in order. You have to use the hover to determine the reason why a document was sorted a particular way.

Depending on the state of the item in the predictive model, the Prediction column can display different types of information:

- Ranking Bar— Gives tag and prediction rank for the parent item. You can sort on the column to display in descending or ascending order. The sorting is based on the highest Prediction Rank in the entire document family. See *"To read Prediction results (Ranking Bar example)" on page 66*.
- Training—Indicates the item was used for training.
- Unknown—Indicates no Prediction Rank for the item due to content characteristics that are not known to the system. The unknown state can change with additional training cycles.
- No Indexable Content—Indicates no Prediction Rank because the item has no text content.

Reading Prediction Rank Results

Information: The Prediction column:

- Appears only if Transparent Predictive Coding actions have been run and you have View access.
- Remembers and displays the previously viewed Predictive Tag.
- Displays other conditions such as Training, Unknown and No Content, when appropriate, instead of the bar chart.

User: Case Administrator, Case Managers and Review Team

To read Prediction results (Ranking Bar example)

1. Select the predictive tag from under the edit icon.
2. Hover over the prediction bar to get additional predictive tag information.

   Note: Sorting on the Prediction column uses the document family score and not the score for the email item itself. This means that when you sort, the Prediction rank bars may not be in order. You should use the hover action to determine the reason why a document was sorted a particular way.

   This example uses the Responsive tag. The Prediction rank displays a ranking of 87.09% for the parent item and also gives you the (38.7%) prediction ranking of the attachment item.

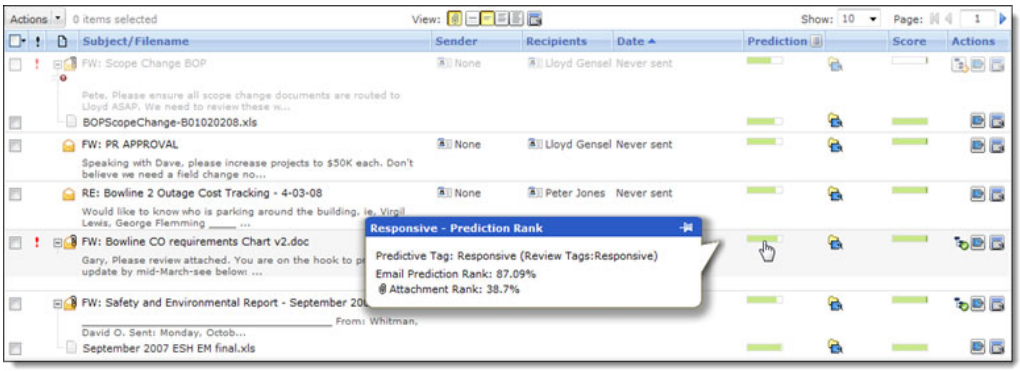


Table 20: Search Prediction Rank Information

Field	Description
Predictive Tag Name	Name of the tag
Parent Prediction Rank	Percentage that the parent matches the predictive tag
Attachment or Embedding Prediction Rank	Percentage ranking of the attachment or embedding in the family

3. Click on the **Prediction** bar and the **View Training Examples** menu displays. A list of the items that were considered the most influencing in the tagging prediction for this item displays. For more information, see *"Understanding Predictive Ranking Decisions" on page 73*.

Viewing Detail Prediction Rank Results for Attachments and Embeddings

While the main search page listing gives you a prediction rank for the parent item, you can view prediction ranks for all the related attachments of an item by selecting the **Detail** mode for an item.

To view detailed Prediction Rank results for attachments and embeddings

1. Click on **Detail** for the item you want to view.

   The **Detail** screen opens.

   You can see the prediction ranking for each attachment.

2. Click on the prediction ranking bar for the attachment and the **View Training Examples** menu opens with more detail. See *"Understanding Predictive Ranking Decisions" on page 73*.

Filtering Prediction Rank Results

You can apply different Prediction Rank filters to gain a better understanding of how the Transparent Prediction Coding model is working and intelligently cull down your review population. The platform gives you a ranking of the prediction tags in the order of most to least relevant. For general filtering information, see *"Filtering Search Results" in the Veritas eDiscovery Platform User's Guide*.

Prediction Rank filtering can be valuable in whittling down and prioritizing a subset of items that will ultimately lead to a reduced set of key items for the case. For example, you may want to manage the large item review for responsive documents by assigning review of certain percentage ranges to your reviewers for culling. These filtering ranges can be useful as a First-Pass Review in order to get a quick initial batch of relevant review items for your review team.

To filter on Prediction Rank Results

1. Click on the filter category to sort the items.

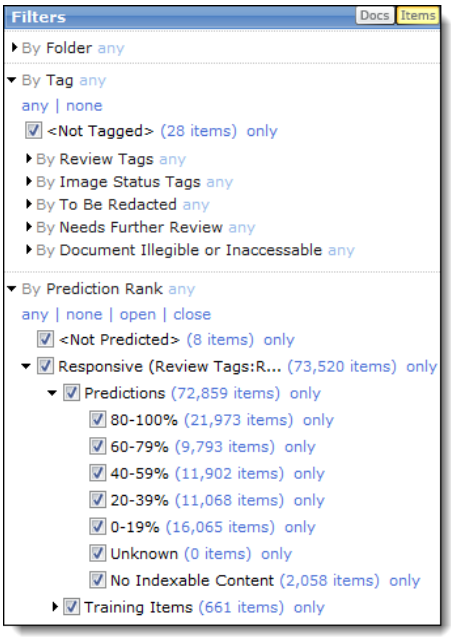


Table 21: Filter Search Results: Prediction Rank

Filter Name	Usage Guidelines
Prediction Rank	Includes all prediction tags. It also includes 2 subfilters:
• Predictions subfilter— Lists all the items that have been run through the Predict action.	– 5 predefined ranges of prediction ranks. – No Indexable Content — Includes items that could not be assigned a Prediction Rank. – Unknown — Includes all items that do not match the training set.
• Training subfilter — Lists all items that have run through the Train action.	– Training — Includes all items used for training the predictive tag. – Not Used — Includes all items that were not used for training due to training errors. **Note:** The search training filter separates items into Trained and Not Used whereas Advanced Search Training Document combines both of these elements into Training Document.

Prediction Searches

Find particular prediction rankings and ranges on your items with Prediction search to gain insight into your data. For example, depending on the review requirements of the case, you can use Prediction search to locate all the Responsive items between a certain percentage range and exclude all Privileged items in your search.

You perform a Prediction Search from the Advanced Search page. The Prediction search can be expanded using multiple rows, each of which contains three drop-down boxes and a percentage min and max field.

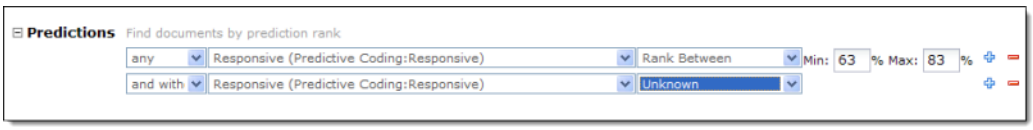


Table 22: Advanced Search: Predictions

Field/Option	Usage Guidelines
any/not any with/without	Enables positive or negative searches
Tag name	Displays a list of all predictive tags

Table 22: Advanced Search: Predictions

Field/Option	Usage Guidelines
Rank Between/ Unknown/ No Indexable Content/ Training Document'	Finds items according to the following rules: • **Ranked Between** specifies the range of Prediction Ranks • **Unknown** includes all items that do not match the training set. • **No Indexable Content** includes items that could not be assigned a Prediction Rank • **Training Document** includes all items used for training the predictive tag
All Control Set	Includes all items selected for the Initial Sample as well as the Additional Sample for the Predictive Tag
Only Initial Sample	Includes only items selected for the Initial Sample for the Predictive Tag
Only Additional Sample	Includes only items selected for the Additional Sample for the Predictive Tag
Min%/ Max%	Specifies the minimum and maximum prediction ranking

Viewing Prediction Rankings

Transparent Predictive Coding Platform offers complete visibility into an item's prediction. Not only does the system calculate prediction rankings but it shows why and what the system considered when assigning a particular prediction ranking. Seeing how the system arrived at its results ensures the quality, defensibility, and statistical veracity of those results and validates the training process.

The review platform displays prediction rankings in the tagging panel for both an item and its attachments. In addition, the View Training Examples screen along with the Analyze functionality provide in-depth analysis of why the model applies a tagging decision for a given prediction ranking. This transparency provides assurances into the reliability and enables you to make informed decisions about your review strategy and scope.

Refer to the following Review topics:

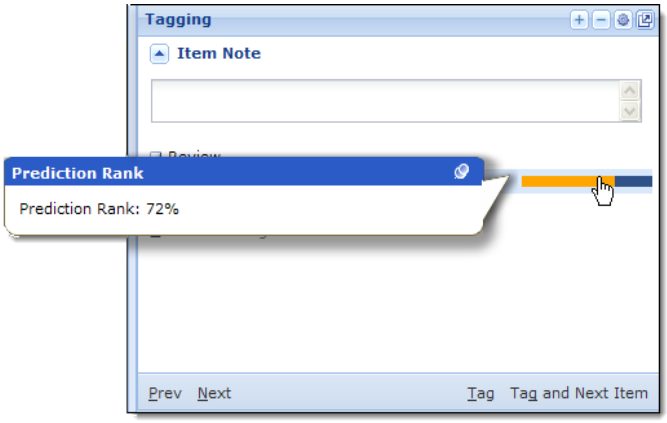
- *"Viewing Tagging Prediction Rank: Review Mode" on page 71*
- *"Viewing Attachment Prediction Rank: Review Mode" on page 72*
- *"Viewing Predictive Tagging events: Review Mode" on page 73*
- *"Understanding Predictive Ranking Decisions" on page 73*

Viewing Tagging Prediction Rank: Review Mode

You can view the prediction rank of every predictive tag in the tagging panel.

To view prediction ranking in the tagging panel

1. In review mode, select the tag and hover over it to see its prediction rank.



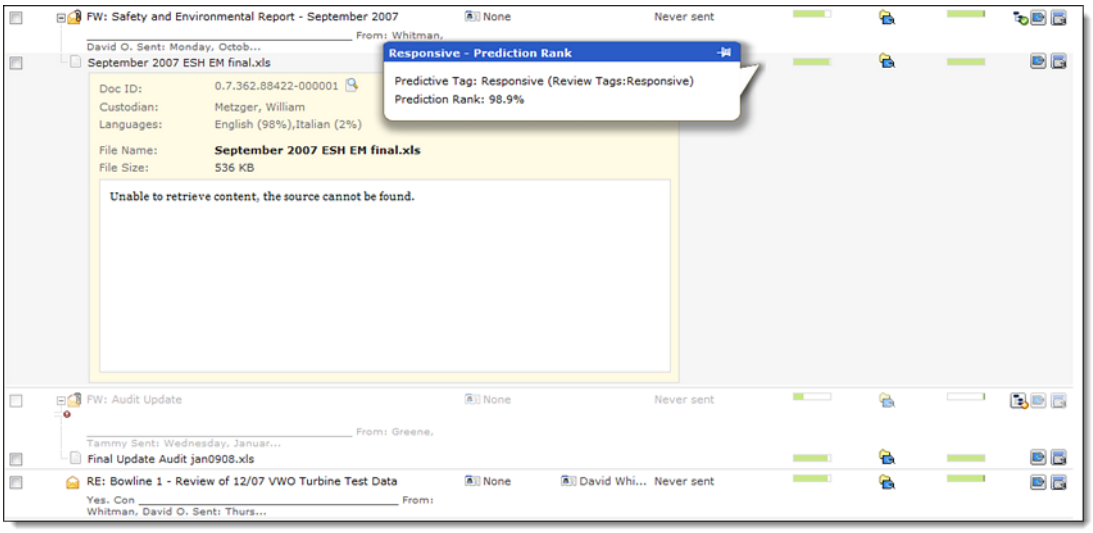
2. To open the View Training Examples menu, click on the prediction ranking bar. For more information, see *"To view the View Training Examples" on page 73*.

Viewing Attachment Prediction Rank: Review Mode

You can view the prediction ranks for all the attachments of the item you are reviewing. In this example, there are 8 attachments with prediction rankings for the Responsive tag and it shows the prediction rank for the third attachment.

To view prediction ranks of attachments

1. In review mode, hover over the prediction bar to get the prediction rank for the attachment.

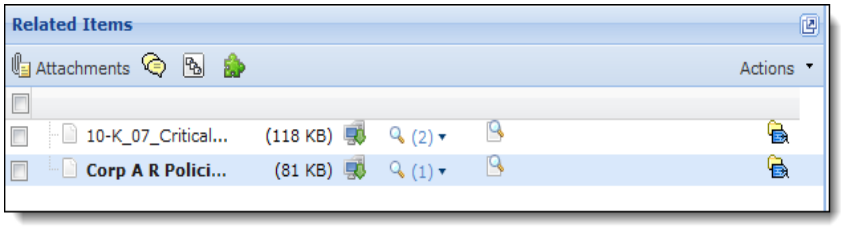


2. To open the **View Training Examples** menu, click on the prediction ranking bar for the attachment. See *"Understanding Predictive Ranking Decisions" on page 73*.

Viewing Predictive Tagging events: Review Mode

Use the Related Item**s** pane to view prediction tag event history.

1. The Related Items view, which is part of the Review mode, displays prediction rankings and other event history.



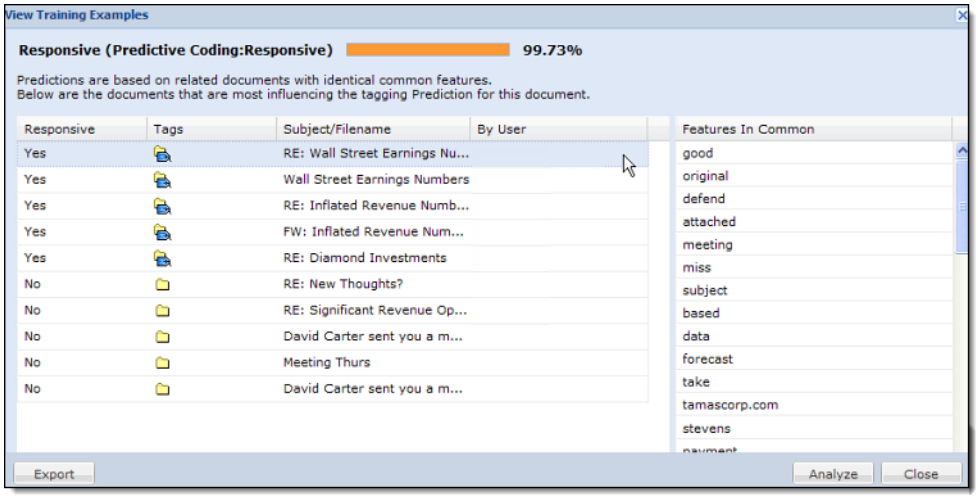
Understanding Predictive Ranking Decisions

The View Training Examples screen shows you how the system derived its prediction ranking. Use the table entries to understand which training items have the highest influence on the tagging prediction for the item. From this screen, you can export the contents to a CSV file or select **Analyze** to drill down to the Predictive Tag Analysis menu for further analysis and, if necessary, perform actions to retrain the system.

To view the View Training Examples

1. Click on the prediction bar ranking from either the Search or Review screen display of the item.

   The View Training Examples screen displays.



2. Clicking on an item displays all the entries tagged with the Predictive Tag (for example Responsive) or close matches of those that were under the **Features In Common** heading.

3. Click **Export** to copy the contents to a CSV file.

4. Click **Analyze** to open the **Predictive Tag Analysis** menu. See *"Predictive Tag Analysis" on page 75*.

Predictive Tag Analysis

You can glean deeper insight, adjust and correct tagging decisions from the Predictive Tag Analysis screen. The transparency of the system enables you to evaluate its approach by looking at the primary item, positive and negative examples, and the highlighted **Features in common** elements. This information helps you to determine if further refining is necessary. For example, if you decide that a certain item should be tagged differently by the system, you can retag, retrain and then predict with the new information. You can use the **Actions** menu for the retraining workflow of: **Tag**, **Train** and **Predict**.

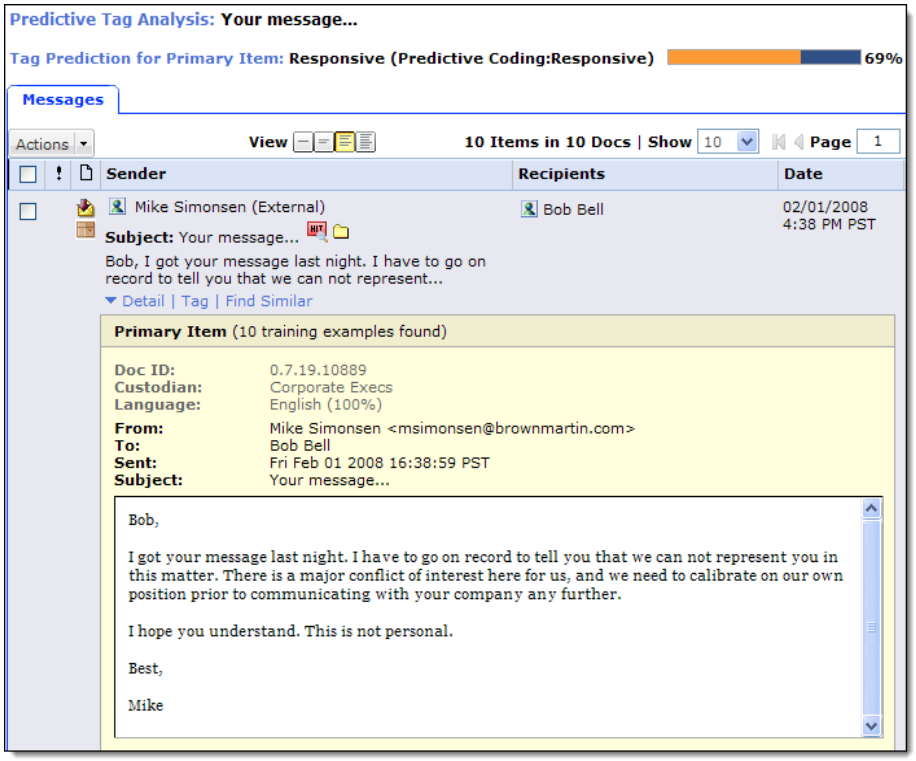
Important: If you do change the tagging, make sure that you immediately incorporate the changes into the system by cycling through the Train and Predict workflow. If you do not retrain the model with the new tagging decisions, the changes have no effect until the system is made aware of the changes during training.

To analyze predictive tag choices

The following example shows the primary item along with negative and positive examples that account for the predictive tagging decision of Responsive. Examining both the negative and positive examples gives you insight as to why the system applied what it learned and how useful it is in its current state for predictive tagging.

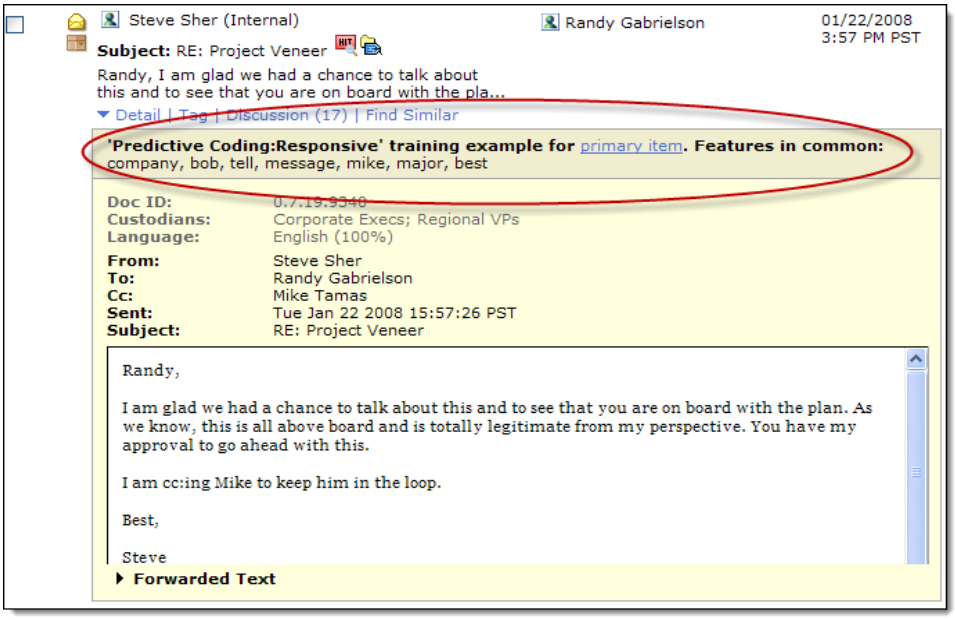
- From the **View Training Examples** screen, click **Analyze**.

  The Predictive Tag Analysis Screen displays.



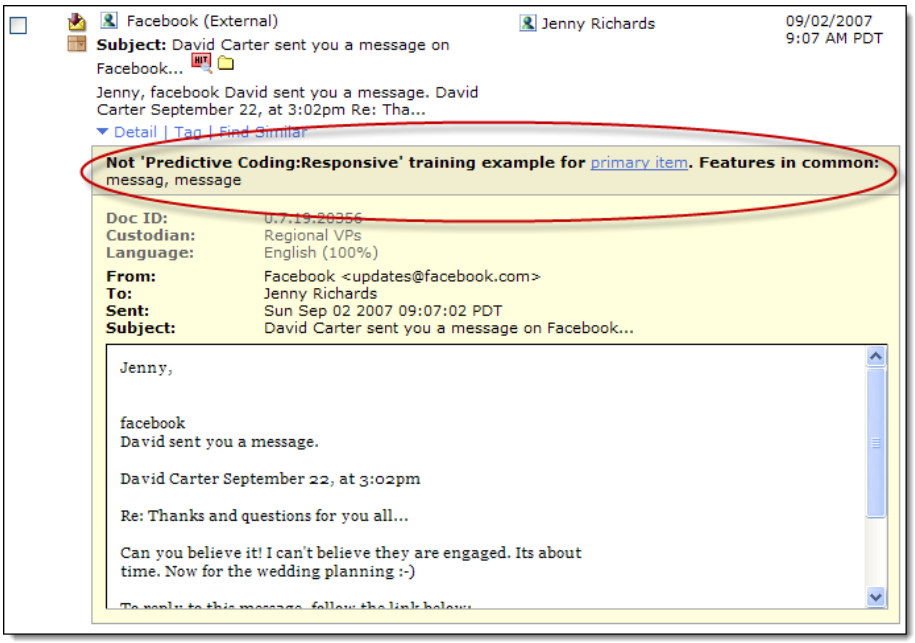
  The Primary Item displays first.

– This is a positive example.



The heading indicates this is a positive example and lists the associated **Features in common**.

– This is a negative example.



The heading indicates it is a negative example and lists the **Features in common**.

Appendix A: Review Accuracy and Metrics

This appendix covers the following topics related to prediction accuracy testing:

Tests

Controlled Prediction Accuracy Test and Prediction Accuracy Test

How do these tests work?

The tests compare the system's predictive tagging decisions against the humans' tagging decisions and compute the number of True Positives, False Positives, True Negatives, and False Negatives. These metrics factor into the F-Measures and the accuracy of the predictions. The F-Measures calculated by the test can help you determine whether to continue iterating or to move items outside of the Transparent Predictive Coding workflow.

How do you view what decisions the system will make?

To see what decisions the system would make, you should select a Prediction Rank Threshold. At this point, you can see exactly which items would be considered to be positive, (the remainder of the items in the Population would be considered to be negative), by running an advanced search on all items in the Population that have Prediction Ranks at or above the Prediction Rank Threshold.

Which Prediction Rank Threshold should I choose?

For the most balanced results, choose the Prediction Rank Threshold that specified in the test result as the *best* threshold." The *best* Prediction Rank Threshold will coincide with the highest possible F-Measure value calculated by the test, which is available in the table of the test report. The test report also allows you to see how changing the Prediction Rank Threshold affects the F-Measure, Precision, and Recall values. For example, a lower Prediction Rank Threshold may give a higher Recall value with a lower Precision and F-Measure, while a higher Prediction Rank Threshold may give a higher Precision value with a lower Recall and F-Measure.

How does the Controlled Prediction Accuracy Test differ from the Prediction Accuracy Test?

The Controlled Prediction Accuracy Test operates on a system-created Control Set, and contains detailed information about the nature of the Population and the statistics involved with the creation of the Control Set. This means that the generated test results not only contain calculated values of the F-Measure, Precision, and Recall based on the Control Set, but they also contain projected minimum and maximum values for these metrics and are statistically guaranteed for the Population.

The Prediction Accuracy Test, on the other hand, simply operates on a set of items with Prediction Ranks that are specified by the user. Thus, the generated test results only contain calculated values of the F-Measure, Precision, and Recall based on the user-specified items.

Appendix B: Product Documentation

The table below lists the administrator and end-user documentation that is available for the Veritas eDiscovery Platform product.

Veritas eDiscovery Platform Documentation

Document	Comments
Installation and Configuration	
Installation Guide	Describes prerequisites, and how to perform a full install of the Veritas eDiscovery Platform application
Upgrade Overview Guide	Provides critical upgrade information, by version, useful prior to upgrading an appliance to the current product release
Upgrade Guide	Describes prerequisites and upgrade information for the current customers with a previous version of the software application
Utility Node Guide	For customers using utility nodes, describes how to install and configure appliances as utility nodes for use with an existing software setup
Distributed Architecture Deployment Guide	Provides installation and configuration information for the Review and Processing Scalability feature in a distributed architecture deployment
Getting Started	
Navigation Reference Card	Provides a mapping of review changes from 10.x compared to 9.x, 8.x compared to 7.x and 7.x compared to 6.x
Administrator's QuickStart Guide	Describes basic appliance and case configuration
Reviewer's QuickStart Guide	A reviewer's reference to using the Analysis & Review module
Tagging Reference Card	Describes how tag sets and filter type impact filter counts
User and Administration	
Legal Hold User Guide	Describes how to set up and configure appliance for Legal Holds, and use the Legal Hold module as an administrator
Identification and Collection Guide	Describes how to prepare and collect data for processing, using the Identification and Collection module
Case Administration Guide	Describes case setup, processing, and management, plus pre-processing navigation, tips, and recommendations. Includes processing exceptions reference and associated reports, plus file handling information for multiple languages, and supported file types and file type mapping
System Administration Guide	Includes system backup, restore, and support features, configuration, and anti-virus scanning guidelines for use with Veritas eDiscovery Platform
Load File Import Guide	Describes how to import load file sources into Veritas eDiscovery Platform
User Guide	Describes how to perform searches, analysis, and review, including detailed information and syntax examples for performing advanced searches

Veritas eDiscovery Platform Documentation

Document	Comments
Imaging Tool Upgrade Guide	Provides details about the Imaging Tool Upgrade feature and how to perform Imaging Tool Upgrade after the eDiscovery Platform appliance is upgraded to version 10.1, workflows affected when the cases are upgraded or not upgraded, and frequently asked questions (FAQs).
Export and Production Guide	Describes how to use and produce exports, productions, and logs (privilege and redaction logs)
Transparent Predictive Coding User Guide	Describes how to use the Transparent Predictive Coding feature to train the system to predict results from control data and tag settings
Audio Search Guide	Describes how to use the Audio Search feature to process, analyze, search and export search media content
Reference and Support	
Audio Processing	A quick reference card for processing multimedia sources
Audio Search	A quick reference card for performing multimedia search tasks
Legal Hold	A quick reference card of how to create and manage holds and notifications
Collection	A quick reference card of how to collect data
OnSite Collection	A quick reference for performing OnSite collection tasks
Review and Redaction	Reviewer's reference card of all redaction functions
Keyboard Shortcuts	A quick reference card listing all supported shortcuts
Production	Administrator's reference card for production exports
User Rights Management	A quick reference card for managing user accounts
Online Help	
Includes all the above documentation (excluding Installation and Configuration) to enable search across all topics. To access this information from within the user interface, click **Help**.	
Release	
Release Notes	Provides latest updated information specific to the current product release

Glossary Terms

F-Measure

The harmonic mean between Precision and Recall. As Precision and Recall are often inversely correlated (such as Precision increases and Recall decreases, and vice versa), the F-Measure is useful in determining the balance between the two metrics.

False Negative

This occurs when an item's Prediction Rank is below the Prediction Rank Threshold but the item has been tagged as positive.

False Positive

This occurs when an item's Prediction Rank is at or above the Prediction Rank Threshold but the item has not been tagged as positive.

Precision

The ratio of True Positives to the total number of Positive items identified by the system. A higher precision indicates a smaller number of False Positives, while a lower precision indicates a larger number of False Positives.

Prediction Rank

This refers to how closely a given item's content and metadata match what the system deems to be Positive.

Prediction Rank Threshold

This refers to a Prediction Rank value selected by using a test report. All items that have a Prediction Rank at or above the Prediction Rank Threshold would then be considered as positive, while all items that have a Prediction Rank below the Prediction Rank Threshold would then be considered as negative.

Population

All of the items to which Transparent Predictive Coding is being applied.

Recall

This is the ratio of True Positives to the total number of Positive items identified by the system. A higher recall indicates a smaller number of False Negatives, while a lower precision indicates a larger number of False Negatives.

Training Conflicts

For a given set of items for training, Transparent Predictive Coding identifies tagging decisions that mark similar items differently as conflicts.

Note: Including items in training that are marked as conflicts without resolving the underlying reason for the conflict can affect and potentially lower the prediction accuracy.

True Negative

This occurs when an item's Prediction Rank is below the Prediction Rank Threshold and the item has not been tagged as positive.

True Positive

This occurs when an item's Prediction Rank is at or above the Prediction Rank Threshold and the item has been tagged as positive.