Getting Started with Amazon Comprehend: Custom Classification

**SPL-TF-100-MLACCL-1 - Version 1.0.5**

© 2023 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. All trademarks are the property of their owners.

Note: Do not include any personal, identifying, or confidential information into the lab environment. Information entered may be visible to others.

Corrections, feedback, or other questions? Contact us at [*AWS Training and Certification*](https://support.aws.amazon.com/#/contacts/aws-training).

**Lab overview**

Your company, AnyCompany Consulting, wants to integrate natural language processing (NLP) to extract key phrases, entities, and sentiment from documents. They have chosen Amazon Comprehend as their NLP service so they can integrate NLP into their existing applications. To start the process of creating models and endpoints for their documents, they want you to create a custom classification model to identify documents with categories and labels.

In this lab, you create and test a custom classification model using Amazon Comprehend.

OBJECTIVES

By the end of this lab, you will be able to:

* Create a custom classification model using Amazon Comprehend.
* Conduct a real-time analysis with a custom classification endpoint.

TECHNICAL KNOWLEDGE PREREQUISITES

To successfully complete this lab, you should be familiar with basic navigation of the AWS Management console and have knowledge of Amazon S3.

DURATION

This lab requires approximately *75* minutes to complete.

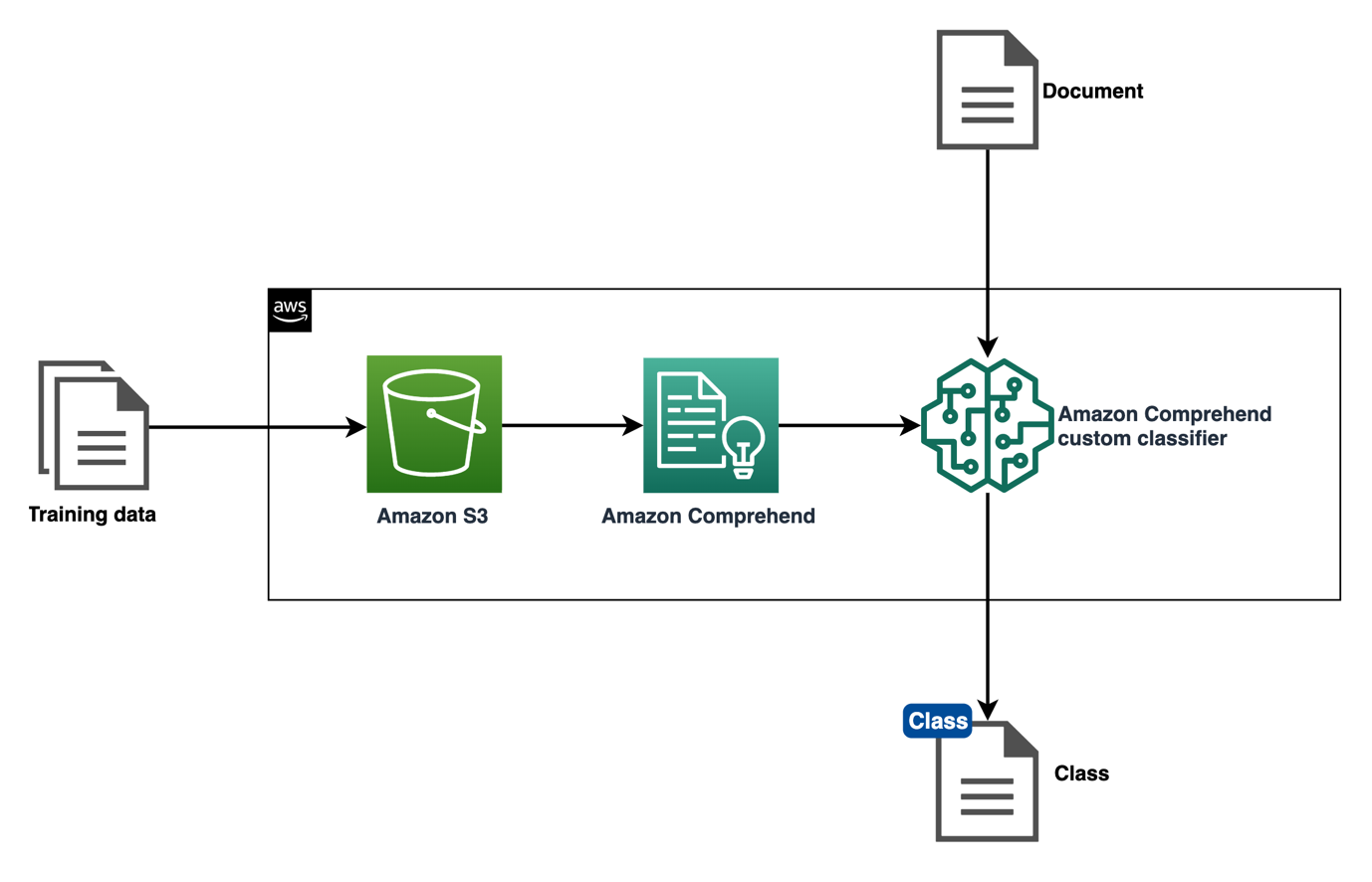
ICON KEY

Various icons are used throughout this lab to call attention to different types of instructions and notes. The following list explains the purpose for each icon:

* **Note:** A hint, tip, or important guidance.
* **Learn more:** Where to find more information.
* **Caution:** Information of special interest or importance (not so important to cause problems with the equipment or data if you miss it, but it could result in the need to repeat certain steps).
* **Consider:** A moment to pause to consider how you might apply a concept in your own environment or to initiate a conversation about the topic at hand.
* **File contents:** A code block that displays the contents of a script or file you need to run that has been pre-created for you.
* **Hint:** A hint to a question or challenge.
* **Answer:** An answer to a question or challenge.

ENVIRONMENT OVERVIEW

The following diagram shows the basic architecture of the lab environment:



*In the preceding diagram, training data is uploaded to an Amazon S3 bucket. Amazon Comprehend uses the training data in the bucket to train a custom classification model. When documents are analyzed in real-time using an endpoint, the trained model classifies the document.*

**Start lab**

1. To launch the lab, at the top of the page, choose **Start lab**.

 You must wait for the provisioned AWS services to be ready before you can continue.

1. To open the lab, choose **Open Console**.

You are automatically signed in to the AWS Management Console in a new web browser tab.

**Do not change the Region unless instructed.**

COMMON SIGN-IN ERRORS

**Error: You must first sign out**



If you see the message, **You must first log out before logging into a different AWS account:**

* Choose the **click here** link.
* Close your **Amazon Web Services Sign In** web browser tab and return to your initial lab page.
* Choose **Open Console** again.

**Error: Choosing Start Lab has no effect**

In some cases, certain pop-up or script blocker web browser extensions might prevent the **Start Lab** button from working as intended. If you experience an issue starting the lab:

* Add the lab domain name to your pop-up or script blocker’s allow list or turn it off.
* Refresh the page and try again.

**Task 1: Review the document in the S3 bucket**

AnyCompany Consulting has given you a set of news data organized into categories (classes) to train your first custom classification model using Amazon Comprehend. Review the document before you create your model.

1. At the top of the AWS Management Console, in the search bar, search for and choose

S3

.

1. Choose the link for the bucket name that starts with **databucket**.

There is one CSV file in the bucket named **custom\_news\_classification.csv** that contains labeled news data. There are 1000 records in the dataset that are treated as individual documents during model training.

**File contents:** The CSV file contains data similar to this:

|  |  |
| --- | --- |
| WORLD | PATTANI, Thailand, December 5 (IslamOnline.net amp; News Agencies) - A fleet of military and civil aircraft and helicopters dropped Sunday, December 5, some 120 million paper birds across Thailands troubled Muslim-majority south, described by activists as |
| SCI\_TECH | America Online and WebEx Communications are evaluating ways to adapt some of the online meeting services they currently offer to business users of AOL #39;s AIM instant messaging network to make these services appealing to other AIM consumers. |
| BUSINESS | Motorists coasted through Pennsylvania Turnpike toll plazas for free on Wednesday morning after roughly 1,300 unionized toll collectors and maintenance workers went on |
| SPORTS | It no longer matters that Montana #39;s drive to the NCAA Division I-AA football championship game has featured six straight homefield victories - all |
| WORLD | For years, Israel has feuded with the United Nations refugee agency for Palestinians over a wide range of issues, and recently Israel thought it had found a smoking gun to press its case. |
| BUSINESS | DaimlerChrysler AG reported third quarter profits of almost 1 billion euros on Thursday, rebounding from a year-earlier loss, thanks to a strong performance from a Chrysler division rejuvenated by new models like the 300C sedan and Dodge Magnum wagon. |
| SPORTS | Chicago Cubs general manager Jim Hendry said on Tuesday that he #39;s hoping time will heal some of the wounds created by Sammy Sosa #39;s early exit from the season finale and |
| WORLD | Saudi officials say security forces have killed a suspected militant in the western city of Jeddah after the man tried to use a hand grenade against them. |
| WORLD | BAGHDAD, Iraq - Iraqi militants said they shot and killed an Italian citizen after he tried to break through a guerrilla roadblock on a highway outside the insurgent stronghold of Ramadi. |
| SPORTS | The latest personality to pile into the situation concerning the future, if any, of the British Grand Prix is former World Champion Nigel Mansell. |

**Consider:** How many classes are included in this dataset?

There are four classes included in the dataset: **WORLD**, **SPORTS**, **BUSINESS**, and **SCI-TECH**.

**Note:** If you want to download and view a file, select the filename and choose **Download**.

 Congratulations! You have successfully reviewed the document in the S3 bucket.

**Task 2: Create a custom classification model**

AnyCompany wants a model that can classify news headlines by different news classes. In this task, you create a custom classification model using the news dataset that classifies the beginning of any news article into one of the four classes: WORLD, SPORTS, BUSINESS, or SCI-TECH using Amazon Comprehend.

1. At the top of the AWS Management Console, in the search bar, search for and choose

Amazon Comprehend

.

1. Choose **Launch Amazon Comprehend**.
2. In the navigation pane at the left of the page, in the **Customization** section, choose **Custom classification**.
3. Choose **Create new model**.
4. On the **Create new model** page, in the **Model settings** section:

* For **Model name**, enter

news

.

* For **Version name**, enter

1

.

1. In the **Data specifications** section:

* For **Classifier mode**, choose **Using Multi-label mode**.
* For **Training data location on S3**, copy and paste the **TrainingDataLocation** value that is listed to the left of these instructions.

News articles might belong to more than one class, so your model uses multi-label mode to reflect that possibility.

1. In the **IAM role** section:

* For **IAM role**, choose **Use an existing IAM role**.
* For **Role name**, choose **ComprehendClassificationRole**.

1. At the bottom of the page, choose **Create**.

In the **Classifier models** section, a **news** model appears with a **Version status** of **Submitted**.

**Note:** This model takes 35-45 minutes to train. While you wait for the model to train, explore real-time analysis and the Amazon Comprehend documentation in the next task.

**Learn more:** Refer to [Custom classification](https://docs.aws.amazon.com/comprehend/latest/dg/how-document-classification.html) for more information about how custom classification works in Amazon Comprehend and the options for creating a custom model.

 Congratulations! You have successfully created a custom classification model.

**Task 3: Explore real-time analysis**

Amazon Comprehend can use built-in or custom models to analyze text in real-time. You can recognize entities, extract key phrases, detect dominant languages, detect Personally Identifiable Information (PII), determine sentiment, detect targeted sentiment, or analyze syntax. You can also use custom models to detect entities or classify documents.

In this task, you use real-time analysis to explore Amazon Comprehend in more depth.

1. In the navigation pane at the left of the page, choose **Real-time analysis**.
2. In the **Input data** section, delete the text and enter:

Hello Zhang Wei, I am John. Your AnyCompany Financial Services, LLC credit card account 1111-0000-1111-0008 has a minimum payment of $24.53 that is due by July 31st. Based on your autopay settings, we will withdraw your payment on the due date from your bank account number XXXXXX1111 with the routing number XXXXX0000.

Customer feedback for Sunshine Spa, 123 Main St, Anywhere. Send comments to Alice at sunspa@mail.com.

I enjoyed visiting the spa. It was very comfortable but it was also very expensive. The amenities were ok but the service made the spa a great experience.

1. Choose **Analyze**.

The results appear in the **Insights** section.

In the **Insights** section, you can choose any of the tabs to view different analyses. The first tab is **Entities**.

The **Entities** tab lists each entity, and the level of confidence that Amazon Comprehend has detected in the input text.

**Consider:** How many types are shown in the list of entities?

Amazon Comprehend identified 12 entities with 6 types in the sample data: **Person**, **Organization**, **Quantity**, **Date**, **Location**, and **Other**.

1. Choose **Application integration** to view the API call and API response details.

You can see more details about the returned results in the **Application integration** section.

1. Choose **Application integration** to close the API call and API response details.
2. Choose the **Key phrases** tab.

The **Key phrases** tab lists the key noun phrases that Amazon Comprehend detected in the input text and the associated confidence level.

The key phrases are underlined. If you choose any key phrases in the **Analyzed text** section, it displays the **Confidence** value for that phrase.

1. Choose the **Language** tab.

The **Language** tab shows the dominant language of the text and Amazon Comprehend’s level of confidence that it has detected the dominant language correctly.

1. Choose the **PII** tab.

The **PII** tab lists entities in your input text that contain PII. A PII entity is a textual reference to personal data that can be used to identify an individual, such as an address, bank account number, or phone number.

The **PII** tab can analyze offsets or labels. The offsets analysis identifies the location of PII in your document. The labels analysis identifies PII in your document and returns the labels of identified PII entity types.

1. Choose **Labels**.

**Consider:** How many PII labels are found? Are there any other types of PII you want to check for?

Amazon Comprehend identified 5 PII labels: **Email**, **Name**, **Bank account number**, **Bank routing**, and **Credit debit number**.

**Learn more:** Refer to [Detecting PII entities](https://docs.aws.amazon.com/comprehend/latest/dg/how-pii.html) for more information about the PII entities Amazon Comprehend can identify.

Now, add more PII to the text field and run the analysis again.

1. In the **Input text** section, at the end of the paragraph, enter

Please contact me at 111-222-3334 if you have any questions.

.

1. Choose **Analyze**.

The results appear in the **Insights** section.

1. Choose the **PII** tab.
2. Choose **Labels**.
3. The **Phone** type is identified with a **Confidence** of **1.00**.
4. Choose **Offsets**.

The PII text is underlined. If you choose any PII text in the **Analyzed text** section, it displays the **Entity** and **Confidence** value for that PII text.

1. Choose the **Sentiment** tab.

The **Sentiment** tab shows the dominant sentiment of the text. Sentiment can be neutral, positive, negative, or mixed. In this case, each sentiment has a confidence rating, providing an estimate by Amazon Comprehend for that sentiment being dominant.

1. Choose the **Targeted sentiment** tab.

The **Targeted sentiment** tab shows the sentiments expressed about entities mentioned in the text. Amazon Comprehend assigns a sentiment rating to each mention of an entity, along with a confidence rating and other information.

The text with sentiment is underlined. If you choose any underlined text in the **Analyzed text** section, it displays more details about the entity type, entity confidence, sentiment, sentiment confidence, and total related entities where applicable.

1. In the **Results** table, choose the numbers above the table to view more pages of the table results. Ten results are shown per page.
2. Choose the plus **+** icon to expand any of the related entities groups.

**Consider:** Take a moment to explore the related entity groups. Do any of the related entities stand out?

Amazon Comprehend identified the customer name and the pronouns in the text that refer to the customer. It also distinguished the customer from the AnyCompany Financial customer service representative, John.

1. Choose the **Syntax** tab.

The **Syntax** tab shows a breakdown of each element in the text, along with its part of speech and the associated confidence score.

1. In the **Input data** section, delete the text and enter:

Hello John - Thank you for your interest in signing up for our news service. AnyCompany News uses the latest machine learning technology to find news that you want to read, when you want to read it. We specialize in Sports, Business, World, and Science/Technology news.

If you have any questions about your subscription, please contact us at support@anycompanynews.com or at 1-800-111-2222.

This email was sent to john@anycompanyconsulting.com. If you no longer want to receive this email, you can unsubscribe here.

1. Choose **Analyze**.

The results appear in the **Insights** section.

1. Explore the tabs in the **Insights** section for this new text document.

CHALLENGE A

Take a moment to write your own input data. Analyze the text and see what results are returned.

Expand the **Hint** or **Solution** sections if you want help solving the challenge.

**Hint**

**Solution**

You have conducted a real-time analysis. Your custom classification model training might be complete now.

1. In the navigation pane at the left of the page, in the **Customization** section, choose **Custom classification**.
2. Wait until the **news** model **Version status** changes to **Trained**.

**Learn more:** Refer to [Insights](https://docs.aws.amazon.com/comprehend/latest/dg/concepts-insights.html) for more information.

 Congratulations! You have successfully explored real-time analysis.

**Task 4: Use a custom classification model**

Your model is trained and is ready for real-time analysis. In this task, you view the model performance, create an endpoint for custom classification, and use the model in a real-time analysis.

TASK 4.1: VIEW THE MODEL PERFORMANCE

First, view the **news** model to see how the model training performed.

1. In the navigation pane at the left of the page, in the **Customization** section, choose **Custom classification**.

**Refresh:** To see the current **Version status**, refresh your browser tab.

**Caution:** If the **news** model has not finished training, wait until **Version status** changes to **Trained**.

1. Choose **news**.

**Note:** You can retrain your custom classification model with more documents. If you make a new version of the model, the version appears in this section.

1. Choose **1** to open the model.

The **Version details** section contains information about the model, version, how long the training took, the model mode, the number of labels, the number of trained documents, the number of test documents, and other important model training details.

1. Choose the **Performance** tab.

**Consider:** What are the accuracy, precision, and recall scores for the model?

The model scores range from 0.80-0.85. You can improve these scores with better training data, more data from under-represented labels, or by reducing skew in your data. In this lab, a model with scores in this range is sufficient to identify news headlines.

**Learn more:** The metrics provide an insight into how your custom classifier performs during a classification job. If the metrics are high, it is likely that the classification model is effective for your use case. Refer to [Custom classifier metrics](https://docs.aws.amazon.com/comprehend/latest/dg/cer-doc-class.html) for more information about Amazon Comprehend custom classification model scores.

You have viewed the model performance and are ready to create an endpoint.

TASK 4.2: CREATE AN ENDPOINT

Next, create an endpoint to use in a real-time analysis.

1. Choose the **Endpoints** tab.
2. Choose **Create endpoint**.
3. For **Endpoint name**, enter

news-endpoint

.

1. For **Number of inference units (IUs)**, enter

1

.

1. Choose the **Acknowledge** message to approve the endpoint charges.

**Learn more:** You do not incur separate charges by using the endpoint in this lab. The endpoint gets deleted automatically when the lab ends. If you create an endpoint in your production environment, there are additional charges. Refer to [Amazon Comprehend Pricing](https://aws.amazon.com/comprehend/pricing/) for more information.

1. Choose **Create endpoint**.
2. Wait until the **Status** changes from **Creating** to **Active**.

**Refresh:** To see the current **Status**, refresh your browser tab.

**Note:** This endpoint takes 10-15 minutes to create.

**Learn more:** Refer to [Managing Amazon Comprehend endpoints](https://docs.aws.amazon.com/comprehend/latest/dg/manage-endpoints.html) for more information about using, monitoring, updating, and deleting Amazon Comprehend endpoints.

You have created an endpoint and are ready to use it in a real time analysis.

TASK 4.3: USE THE MODEL IN A REAL-TIME ANALYSIS

Finally, use your *news* custom classification model in a real-time analysis, testing headlines to see which labels the model identifies.

1. In the navigation pane at the left of the page, choose **Real-time analysis**.
2. In the **Input data** section:

* For **Analysis type**, choose **Custom**.
* For **Custom model type**, choose **Custom classification**.
* For **Endpoint**, choose **news-endpoint**.
* For **Input text**, clear the text and enter:

Seattle, Washington - Senior AnyCompany Consulting executive promoted to new executive chair role, replacing Josh Smith. The change indicates that market forces are starting to push for higher earnings over long-term growth.

1. Choose **Analyze**.

The results appear in the **Insights** section.

1. For **Input text**, clear the text and enter:

Tokyo (Reuters) - The political crisis from the most recent demonstrations destabilized the national leader, leading to fractured insurgent movements while a new leader is elected.

1. Choose **Analyze**.

The results appear in the **Insights** section.

CHALLENGE B

Take a moment to write your own news headline. Analyze the text and see what labels the model classifies the document with.

Expand the **Hint** or **Solution** sections if you want help solving the challenge.

**Hint**

**Solution**

 Congratulations! You have successfully used a custom classification model in a real-time analysis.

**Conclusion**

 Congratulations! You now have successfully:

* Created a custom classification model using Amazon Comprehend.
* Conducted a real-time analysis with a custom classification endpoint.

**End lab**

Follow these steps to close the console and end your lab.

1. Return to the **AWS Management Console**.
2. At the upper-right corner of the page, choose **AWSLabsUser**, and then choose **Sign out**.
3. Choose **End lab** and then confirm that you want to end your lab.

For more information about AWS Training and Certification, see [*https://aws.amazon.com/training/*](https://aws.amazon.com/training/).

*Your feedback is welcome and appreciated.*  
If you would like to share any feedback, suggestions, or corrections, please provide the details in our [*AWS Training and Certification Contact Form*](https://support.aws.amazon.com/#/contacts/aws-training).