



AWS Academy Natural Language Processing  
Module 04 Student Guide  
Version 0.1.0  
200-ACMNLP-01-EN-SG

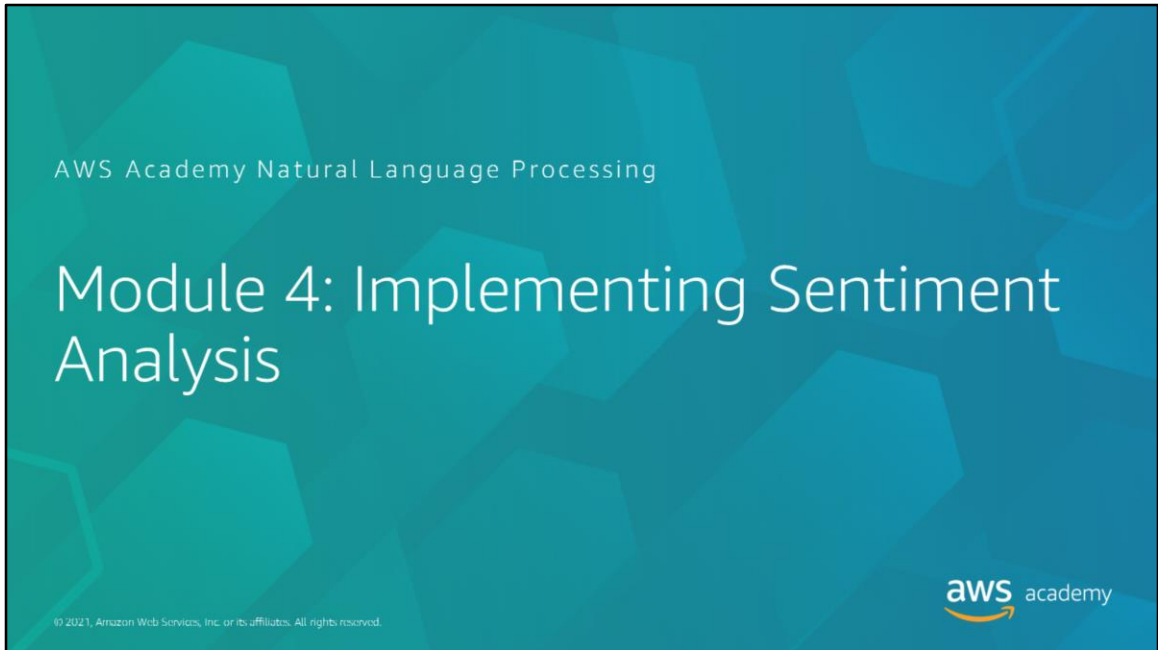
© 2020 Amazon Web Services, Inc. or its affiliates. All rights reserved.

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.


# Contents

Module 4: Implementing Sentiment Analysis	4
---	---



Welcome to Module 4: Implementing Sentiment Analysis.

## Module overview



Sections	Lab
<ul style="list-style-type: none"><li>• Section 1: Introducing the scenario</li><li>• Section 2: Identifying the steps for text processing</li><li>• Section 3: Examining the algorithms for sentiment analysis</li><li>• Section 4: Discussing and walking through the lab solution</li></ul>	<ul style="list-style-type: none"><li>• Challenge Lab: Implementing Sentiment Analysis</li></ul>

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.2

In this module, you will learn about the following topics:

1. Scenario introduction – You learn about the scenario that you will use throughout this module and in the lab.
2. Identifying the processing steps – In this section, you examine the processing steps that are typically used in sentiment analysis, and determine which steps might be useful for the scenario.
3. Examining the algorithms for sentiment analysis – You look at some of the algorithms that you could use to solve this problem. You will also look at what fully managed services can do in this business scenario.
4. Lab solution discussion – After completing the lab, you walk through some of the key observations and learnings from the lab.

## Module objectives



At the end of this module, you should be able to:

- Describe the challenges in working with social media data
- Evaluate machine learning (ML) algorithms and tools that are used in natural language processing (NLP)
- Create a solution to a sentiment analysis business problem

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

5

At the end of this module, you should be able to:

- Describe the challenges in working with social media data
- Evaluate machine learning (ML) algorithms and tools used in natural language processing (NLP) for sentiment analysis
- Create a solution to a sentiment analysis business problem

## Module 4: Implementing Sentiment Analysis

### Section 1: Introducing the scenario

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Introducing Section 1: Introducing the scenario.

## Business case



- Movie review – Use the text of the review to calculate:
  - + (positive sentiment) or - (negative sentiment) ratings
  - Count of positive and negative reviews
- Labeled dataset that contains 50,000 reviews


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

3

To help apply the techniques that you learned in previous modules, you work through a small typical sentiment analysis use case. You play the role of a data scientist on a small development team. The organization that you work for maintains a website of movie reviews. A key customer feature has been identified, which is to provide an overall positive or negative evaluation for a particular movie based on the number of positive and negative reviews. You will develop an ML solution that enables developers to create an inference for a movie review. You must analyze the review and indicate whether it is positive or negative.

To help with this task, you have access to a dataset that contains the raw text of 50,000 movie reviews. These reviews have been labeled as either positive or negative.



Dataset	
	
Text	Sentiment
This is a charming movie starring everyone's favorite cartoon chipmunks. In this feature we follow the band of rodents on an unforgettable balloon race around the world. Although there are lows, including an orphan penguin, all in all it's a great family film.	1
I really should have learned more about this movie before renting it. It was one of those movies where you keep watching it figuring it's got to get better. Then, when it ends, you feel stupid for having wasted precious time in your life that you can never get back. The pictures of the shuttle looks like it was done with a little toy inside of a box and the spacewalking scenes were funny because you could see the strings attached to the space suits. The script was lacking and the car chase scene with the guy bleeding and going unconscious was incredible because he drove better than I could have on one of my best days. All in all, I have seen worse but this sure isn't one I'd recommend or want to remember.	0

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved. b

In the example, the dataset consists of two columns. The first is the **Text** column, which contains the raw text of the movie review. The **Sentiment** column contains a **1** if the review is positive and a **0** if the review is negative. No information is provided about the actual movie in the dataset.

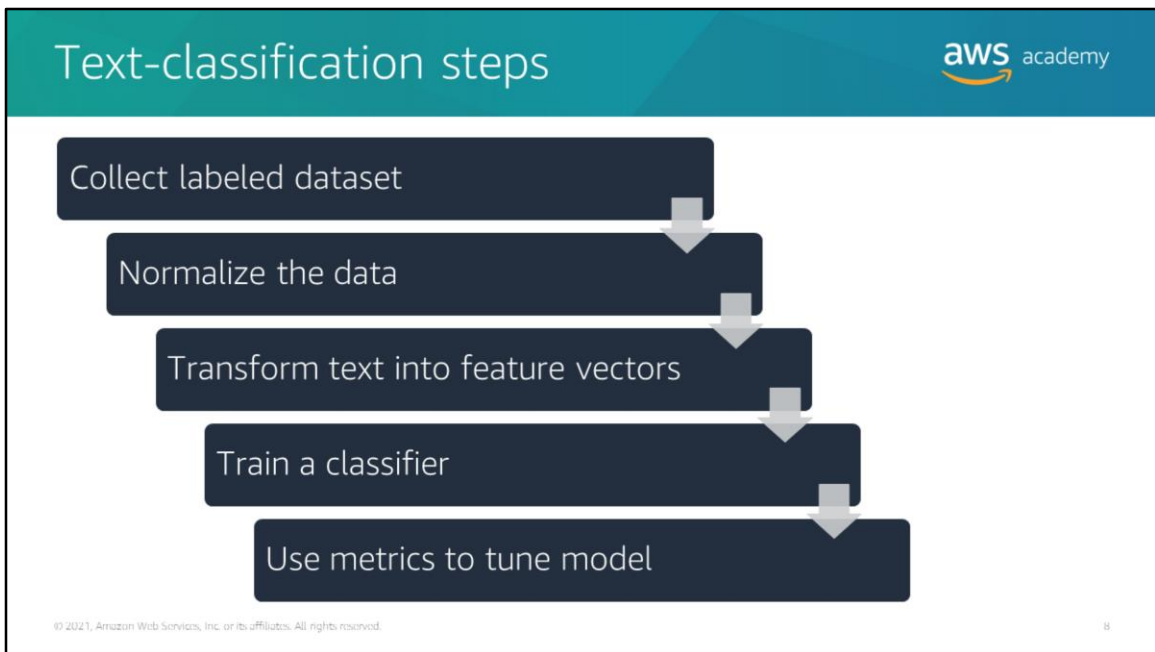
## Module 4: Implementing Sentiment Analysis

### Section 2: Identifying the steps for text processing

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



In this section, you look at what steps you could use when you process text for this business problem.



This business problem seems to be a supervised binary classification problem:

1. You have a labeled dataset. (Supervised)
2. Two values are to be predicted (Positive or Negative)

With the labeled dataset, you decide which normalization tasks are needed to get the **Text** data into a form that's useful to an ML model.

After you normalize and clean the text, you must transform the text into numerical values either as new features or as feature vectors. Whether you choose new features or feature vectors mostly depends on what the algorithm that you select will need.

When the data is ready, you must split it into training and test sets. Depending on the algorithm, you might also need a validation set. You will use this data to train a classifier.

When you have a trained model, it's important that you use metrics to determine how well the model performed. You can try a few different models to find the best one for your data. You might need to tune the models as you proceed.

## What processing steps might you take?



"Diego ""master of debonair"" Ramirez plays the Big Boss (IVAN) who preys upon the unfortunate John Stiles (PINKY) by forcing the hapless ex-con to exploit his ill-found new position in a bank. Nikki Wolf (Miss PELHAM) most effectively provides the female interest, whom Pinkie simply cannot resist.<br /><br />It seems they were unable to decide on one name for this film so instead they used four .... makes sense ???<br /><br />Sadly, this turned out to be one of Niven's last roles.<br /><br />Overall, this film is fun and well worth watching if you manage to catch one of its rare or late night TV screenings.<br /><br />"

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

9

Given the typical example on the slide, what processing steps might you take?

Take a few minutes to think or discuss what steps might be appropriate.

## Handling informal text data



- Informal text data:
  - Markup
  - Non-text data and emojis
  - Shortcuts – 's, 're, 'r
  - URLs
  - Word spacing
  - Non-standard spelling
  - No grammar
  - Multilingual
  - Transliteration
  - Homonyms
- Handling sarcasm, trolls, memes, slang, and fake news

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

10

Movie reviews on a website are usually not written by a trained linguist with 20 years of industry experience. They often contain various informal elements that you must handle. These elements could include:

- Markup (for example, HTML)
- Use of emojis (for example, 😊 😞 😐)
- Abbreviations or shortcuts, such as *'re* for *are* (for example, *we're* instead of *we are*)
- URLs (for example, <http://www.example.com>)
- Word spacing (for example, including more than a single space between words)
- Non-standard spelling—not everyone can spell accurately all the time
- No grammar or punctuation.
- Multilingual—some people use multiple languages when they write on social media or review sites
- Transliteration—when a word from one language is phonetically written in another language, typically using Roman script. For instance, consider the Chinese word 面条. If the word 面条 is transliterated, it would be *mein* (as in the Chinese menu item *lo mein*). *Mein* does not tell you what the original word means in English, but it does help you pronounce it the way that a Chinese speaker would. If you wanted to translate the word, one translation might be *noodles*. For more information about transliteration, see [What is transliteration?](#)
- Homonyms—words with multiple meanings (for example “You are **right**”, “Take a **right** after the bus stop”, and “water is a basic human **right**”. Another example is the use of the word engaged: “The actor wasn’t very **engaged**” and “The protagonist was **engaged** to be married”.)

You also must consider the question of how to handle sarcasm in your review text. In broader

business problems, you might come across trolls, memes, slang, and fake news. Text length could also be an issue with processing social media text, which tends to be shorter than a manuscript or book. For handling special cases—like short social media text—look for a pretrained model that you could use. For example, the Python Natural Language Toolkit (NLTK) has a Tweet tokenizer that can be useful for processing Tweets.

## Processing steps



- Cleaning up text:
  - Convert to lowercase
  - Remove extra spaces
  - Remove HTML tags
- Remove stopwords
  - Exception list
- Tokenize or vectorize text

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

11

The processing steps that seem most likely to be needed are:

- **Cleaning up text** – You might convert all text to lowercase. You could also remove spaces, HTML tags, and other formatting anomalies.
- **Removing stopwords** – You probably want to remove stopwords, but be careful. Removing words that negate statements, such as *not*, can be impactful when you want to predict sentiment.
- **Tokenizing or vectorizing text** – Depending upon which algorithms you use, you can use a simple bag-of-words (BOW) tokenizer, or create vectors by using Word2vec.

## Handling missing data



- Can you fill in missing values?
- If the missing data is a small amount—drop the data
- Use standard missing-data practices for other features

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

12

Handling missing values always provides an interesting set of choices in ML. Generally, if a few text values are missing, you might want to drop the rows altogether. This choice depends on the amount of data that you have. Generally, having no text isn't useful in an NLP problem, unless you have many other features to use. Filling in missing values with similar rows might be risky, but it will depend on the problem that you are trying to solve. For other missing values, a valid choice is to impute the values by using standard ML techniques.



## Section 2: Summary



- ML steps –
  - Gather the dataset
  - Normalize the text
  - Transform the text into features
  - Train a model
  - Use metrics to tune
- Handle informal text data
- Perform processing steps
- Handle missing data

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

15

In this section, you looked at:

- The ML steps – You mapped a typical ML workflow to the business problem, identifying the key steps.
- Handling informal text data – You looked at some of the challenges in handling informal text data, such as reviews or social media.
- Processing steps – You identified the steps that you will take in the lab for this module.
- Handling missing data – You considered how to handle missing text data..

## Module 4: Implementing Sentiment Analysis

### Section 3: Examining the algorithms for sentiment analysis

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



In this section, you will look at the algorithms for sentiment analysis.

## What kind of ML problem is it?



- Supervised
- Binary classification
- Goals are to maximize correct answers
  - Metrics: Error, accuracy

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

15

In the previous section, you learned that the business problem could be solved as a supervised binary classification problem.

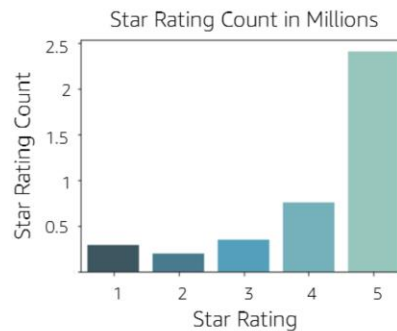
One decision you must make is to decide which metrics you will use to evaluate the model. Because you are concerned with how well both positive and negative reviews are predicted, error or accuracy might be appropriate metrics. What other metrics might you consider?

Whatever you decide, be consistent across different algorithms and services to ensure that you are comparing the performance accurately.

## Class imbalance



- Number of samples per class is **not** equally distributed
- The ML model might not work well for the infrequent classes
- Examples:
  - Fraud detection
  - Anomaly detection
  - Medical diagnosis



[Amazon review dataset](#): The number of five-star reviews almost equals the total of the other four types of star reviews combined


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

1b

One problem that you might encounter is class imbalance, which occurs when the samples you have aren't distributed equally. For example, in fraud detection, you typically have a high percentage of non-fraudulent transactions and only a few fraudulent ones.

The ML model might not work well for infrequent classes. You should consult the documentation of the algorithm to find out whether that is true for the algorithms you are considering. You will look at how to address this situation in the next slide. However, as an example, note that the number of five-star reviews in the [Amazon review dataset](#) is almost equal to the combined total of the other four types of star reviews.

## Addressing the class imbalance problem



Downsampling	Upsampling	Data generation	Sample weights
Reduce the size of the dominant or frequent classes.	Increase the size of the rare or small classes.	Create new records—similar, but not identical. <ul style="list-style-type: none"><li>For example, create similar images that distort original images by rotating, introducing noise, or skewing.</li></ul>	For a model that uses a cost function, assign higher weights to rare classes and lower weights to dominant classes.

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

1/

You can use a few techniques to handle the class-imbalance problem:

- **Downsampling** – Reduce the number of dominant classes so that the mix is more balanced.
- **Upsampling** – Increase the size of the small classes to match the larger class.
- **Data generation** – Use data-generation tools to create new records that are based on the existing ones. A good example for image processing is to rotate the image slightly.
- **Sample weights** – For a model that uses a cost function, assign higher weights to rarer classes.

To make things easier in this module, you will find that the movie review dataset has no class imbalance.

## Classifier performance



- Sparse vectors
  - Vector has a large number of zeros
- Hyperparameter tuning
  - Is important to allocate time for hyperparameter tuning, which can improve the results of a model
  - Understand the metrics and tunable parameters for the algorithm that you are using

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

18

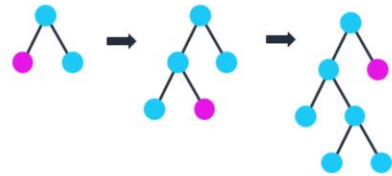
You should keep a couple of other things in mind:

- **Sparse vectors** – When you transform the text data into a numerical representation, you can end up with sparse vectors. A sparse vector has a large number of zeros. You might need to consider how these largely empty vectors are stored and processed. It's often useful to store only the non-zero data, but the tradeoff is that accessing the data becomes more complex. Check with the algorithm that you are using to find out whether sparse vectors could cause issues. XGBoost, which you use in the lab, tends to do well with sparse data—again, to make things easier in the lab.
- **Hyperparameter tuning** – Algorithms can be tuned to get better (or sometimes worse) results. You should make sure that you allocate time to tune your model, which can improve results. Each algorithm will have a set of hyperparameters that can be adjusted. The documentation usually has some tips on which hyperparameters have the most impact and what the ranges are. You create a hyperparameter tuning job to tune the hyperparameters and use the model's metrics to evaluate the best model. You will try this process in the lab.

## XGBoost



- Is an open-source implementation of the gradient boosted trees algorithm
- Has performed well in ML competitions
- Robustly handles various data types, relationships, distributions, and a large number of hyperparameters



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

19

XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable. It attains its prediction by combining an ensemble of estimates from a set of simpler, weaker models. XGBoost has performed well in ML competitions. It robustly handles various data types, relationships, and distributions. It can also handle the many hyperparameters that can be tweaked and tuned for improved fit. This flexibility makes XGBoost a solid choice for problems in regression, classification (binary and multiclass), and ranking.

## BlazingText classification



- BlazingText classification
  - Sentiment analysis
  - Spam detection
  - Hashtag prediction
- Training data format

```
__label__0 this movie is the worst...  
__label__1 One of the best movies..
```
- Recommendations for Amazon Elastic Compute Cloud (Amazon EC2) instances –
  - <2 GB – C5 instance
  - >2 GB – ml.p2 or ml.p3 – Single Graphics Processing Unit (GPU) instances

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.23

BlazingText implements Word2Vec, and it can also perform text classification. This combination works well for sentiment analysis, spam detection, and hashtag prediction. The required data format is straightforward. It requires the label (the 0 or 1 **Sentiment** column) to be preceded by a double underscore, label, double underscore tag. The text should follow this tag, with each record on a single line.

When you train the classifier for smaller datasets (2 GB or less), a C5 Amazon Elastic Compute Cloud (Amazon EC2) instance works well. For larger datasets, try a graphics processing unit (GPU) instance (such as an ml.p2 or ml.p3) for performance.

Note: After the model is trained, it can be run on a smaller instance.



## Using pretrained models



- Natural Language Toolkit (NLTK)
  - *Vader lexicon* – Uses bag-of-words (BOW) approach
- Textblob
  - Uses BOW classifier
  - Includes subjectivity analysis
- Flair
  - Based on character-level, long short-term memory (LSTM) neural network
  - Factors sequences of letters and words
  - Can handle out-of-vocabulary (OOV) words

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

21

In addition to training your own model, you can use a pretrained model. You can start with a pretrained model to see the results that you get.

Natural Language Toolkit (NLTK) contains a bag-of-words (BOW) implementation that is named *Vader lexicon*.

Textblob also uses a BOW classifier, but it also includes subjectivity analysis. Subjective sentences generally refer to personal opinion, emotion, or judgment. In contrast, an objective sentence refers to factual information.

Flair is based on a character-level, long short-term memory (LSTM) neural network. It factors sequences of letters and words, which means that it can handle out-of-vocabulary (OOV) words.

## Amazon Comprehend



- Fully managed service
- Includes application programming interface (API) for sentiment analysis
  - No training required
- Can train a custom classifier
  - Train and test as you would an Amazon SageMaker algorithm
  - Training data must be in correct format
  - Predictions include the class and probability score

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

22

Amazon Comprehend is a fully managed service for finding insights and relationships in text. It includes an application programming interface (API) for sentiment analysis, which can be used without any ML knowledge. You can pass the API some text and receive a response that includes whether the sentiment is positive, negative, mixed, or neutral.

If the default results are not suitable, you can train a custom classifier. You will need both testing and training data for this process (similar to what you would need for XGBoost). You must manipulate the data into the correct format to train the model. After the model is trained, you can perform predictions, which will return the predicted class and the probability score.

## Section 3: Summary



- ML problem
- Class imbalance
- Built-in Amazon SageMaker algorithms
  - XGBoost
  - BlazingText
- Pretrained models
- Amazon Comprehend

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

25

In this section you learned:

- The type of machine learning problem that the movie sentiment analysis problem represents
- How class imbalance could cause issues with certain algorithms, and options for dealing with imbalance
- Two built-in Amazon SageMaker algorithms, XGBoost and BlazingText
- How to use other pretrained models
- How to use Amazon Comprehend

## Module 4 Challenge Lab: Implementing Sentiment Analysis

24



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

You will now complete Module 4 Challenge Lab: Implementing Sentiment Analysis.

## Module 4: Implementing Sentiment Analysis

### Section 4: Discussing and walking through the lab solution

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



In this section, you will walk through the lab solution.

## Lab summary



- Sentiment analysis results

	First pass	Second pass	Third pass	Fourth pass	Fifth pass
Algorithm or service	XGBoost	XGBoost	XGBoost	BlazingText	Amazon Comprehend
Vectorizer	CountVector	CountVector	CountVector	Word2Vec	n/a
Text-processing steps	none	Normalized Stopword removed	Normalized Stopword removed	none	n/a
Hyperparameter tuned (Yes or No)			YES		
~Accuracy	82.22%	51%	85%	87%	88%

Note: Your results will vary from these results.

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

2b

In the lab, you made five passes through the problem by using different normalization techniques, algorithms, and services.

In the first pass, you used a simple CountVector to transform the text into a numerical form for training. Using an untuned XGBoost algorithm, you built a reference to work from.

The second pass introduced some cleanup of the text, in addition to removing most stopwords. You retained some stopwords that might improve the results. You used the calculated metrics to compare this result with the first result. If you had more time, you might have gone through and changed some of the processing steps to see whether the results improved.

The third pass introduced hyperparameter tuning to the XGBoost model. The best model from the job was then used to calculate the metrics, and you again compared these results with the previous results. With more time, you might have changed the parameter tuning, or run more tuning jobs to potentially get better results.

The fourth pass introduced BlazingText, which changed how you processed the text. You needed to do some formatting before you trained the model. For a small amount of work, the results are good.

The final run used Amazon Comprehend, which is a fully managed service. You got acceptable results by using the default model and performing no data processing. In addition, you could complete this run without using a training set.

## Questions?



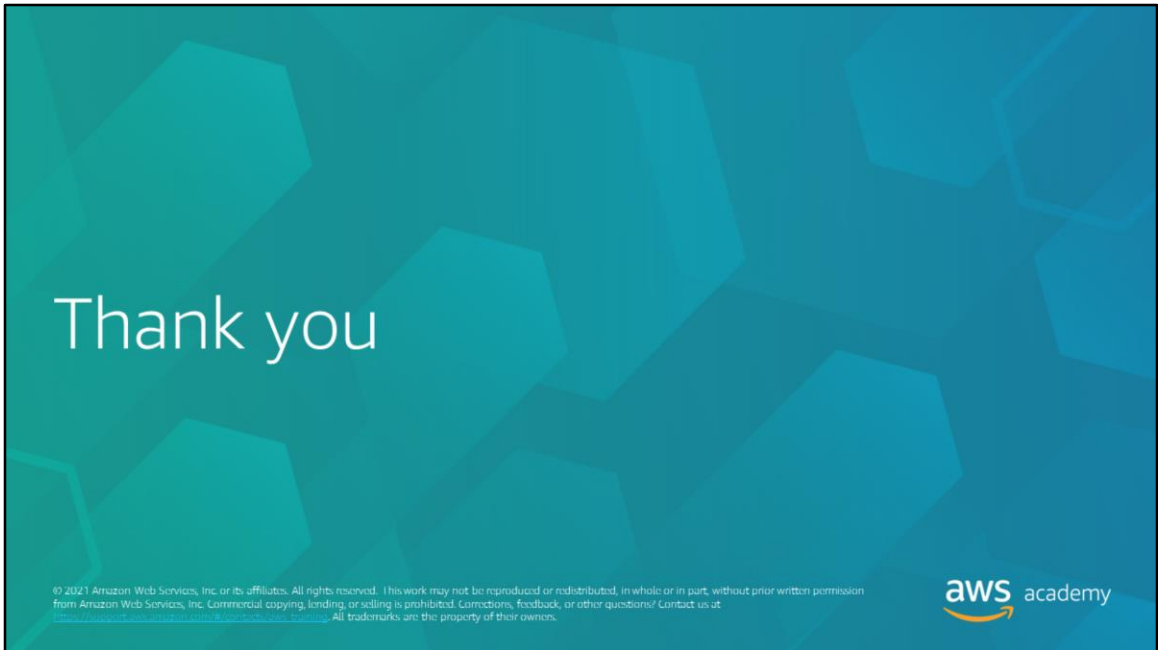
- What might you do differently if you had more time?
- How might you improve the results?

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

2 /

Discuss the questions on the slide with your instructor and classmates.





Thank you for completing this module.