

## MITOCW | MIT15\_071S17\_Session\_5.2.04\_300k

---

In this lecture, we'll be trying to understand sentiment of tweets about the company Apple.

Apple is a computer company known for its laptops, phones, tablets, and personal media players.

While Apple has a large number of fans, they also have a large number of people who don't like their products.

And they have several competitors.

To better understand public perception, Apple wants to monitor how people feel over time and how people receive new announcements.

Our challenge in this lecture is to see if we can correctly classify tweets as being negative, positive, or neither about Apple.

To collect the data needed for this task, we had to perform two steps.

The first was to collect data about tweets from the internet.

Twitter data is publicly available.

And you can collect it through scraping the website or by using a special interface for programmers that Twitter provides called an API.

The sender of the tweet might be useful to predict sentiment.

But we'll ignore it to keep our data anonymized.

So we'll just be using the text of the tweet.

Then we need to construct the outcome variable for these tweets, which means that we have to label them as positive, negative, or neutral sentiment.

We would like to label thousands of tweets.

And we know that two people might disagree over the correct classification of a tweet.

So to do this efficiently, one option is to use the Amazon Mechanical Turk.

So what is the Amazon Mechanical Turk?

It allows people to break tasks down into small components and then enables them to distribute these tasks online

to be solved by people all over the world.

People can sign up to perform the available tasks for a fee.

As the task creator, we pay the workers a fixed amount per completed task.

For example, we might pay \$0.02 for a single classified tweet.

The Amazon Mechanical Turk serves as a broker and takes a small cut of the money.

Many of the tasks on the Mechanical Turk require human intelligence, like classifying the sentiment of a tweet.

But these tasks may be time consuming or require building otherwise unneeded capacity for the creator of the task.

And so it's appealing to outsource the job.

The task that we put on the Amazon Mechanical Turk was to judge the sentiment expressed by the following item toward the software company Apple.

The items we gave them were tweets that we had collected.

The workers could pick from the following options as their response-- strongly negative, negative, neutral, positive, and strongly positive.

We represented each of these outcomes as a number on the scale from negative 2 to 2.

We had five workers label each tweet.

The graph on the right shows the distribution of the number of tweets classified into each of the categories.

We can see here that the majority of tweets were classified as neutral, with a small number classified as strongly negative or strongly positive.

Then, for each tweet, we take the average of the five scores given by the five workers.

For example, the tweet "LOVE U @APPLE" was seen as strongly positive by 4 of the workers and positive by one of the workers.

So it gets a score of 1.8.

The tweet "@apple @twitter Happy Programmers' Day folks!" was seen as slightly positive on average.

And the tweet "So disappointed in @Apple.

Sold me a Macbook Air that WONT run my apps.

So I have to drive hours to return it.

They won't let me ship it." was seen as pretty negative.

So now we have a bunch of tweets that are labeled with their sentiment.

But how do we build independent variables from the text of a tweet to be used to predict the sentiment?

In the next video, we'll discuss a technique called bag of words that transforms text into independent variables.