David Schneider

DA 516, Applications of Data Analytics

**Using Natural Language Processing, Binary Classification, and Deep Neural Networks to Identify Public Transit Adoption of Innovative Mobility Technologies**

**Abstract**

This research uses exploratory data analysis and applies natural language processing, and various binary classifiers to grant application text to identify whether public transportation agencies have adopting innovative mobility technology such as smart-phone apps, electronic fare payment, and transportation coordination software. I trained machine learning algorithms on a data set of 200 labeled records to see how well they predicted which class (conventional or innovative) a technology would fall under. The sci-kit-learn algorithms generally did a poor job of classifying. A Deep Neural Network (DNN) model produced a higher accuracy but also over-fit the training data and I had only limited success to control for over-fitting.

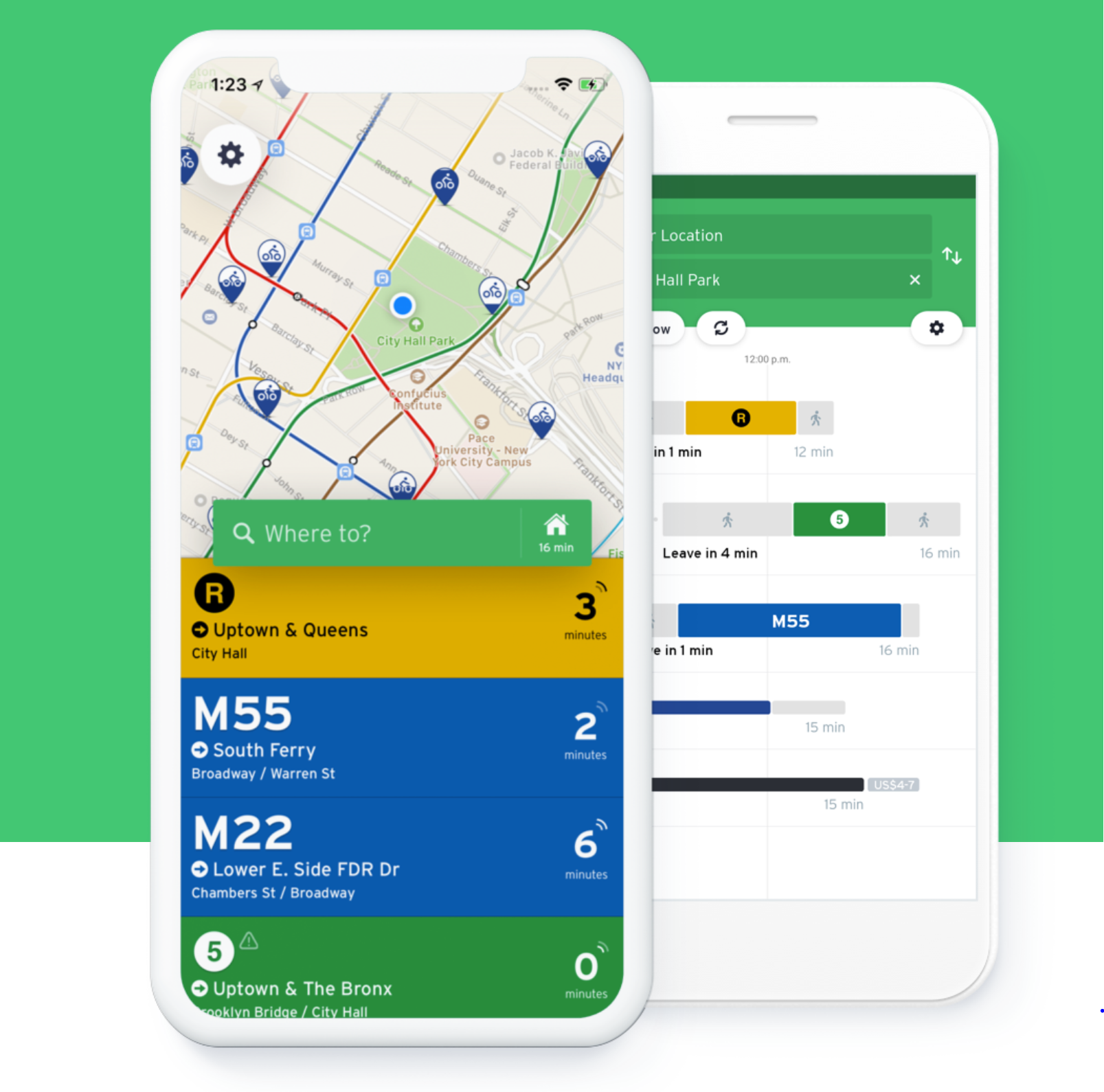
**Background**

Historically, people seeking to use public transportation in the United States have struggled with a service that can be uncertain and unpredictable. Tourists or people new to an area find it difficult to use paper maps or transit agency websites to figure out how to navigate from their origin to destination. Riders approaching a bus or train stop often wonder: when will the next bus arrive? How long will my wait be? Are the buses running on schedule or are they behind or ahead? And, during COVID-19 pandemic, riders may wonder: how crowded is the bus coming up and do I want t get on a crowded vehicle? Transit service can also be slow because people need to pay a fare when they board a bus and, historically, fares have been collected using cash. The time it takes for riders to count out exact change and feed it into a fare box while the bus idles is time that could be spent getting closer to people’s destinations. And then there are people who use public transit that is provided door-to-door (including older adults and people with disabilities who are eligible for paratransit provided by vans because they are not able to use fixed route bus or rail lines). These services are often provided by many different non-profits and social service agencies but, historically, agencies have not coordinated their trip schedules or available seats with one another. This sometimes creates overlapping service and sometimes leaves gaps in service that might otherwise be available.

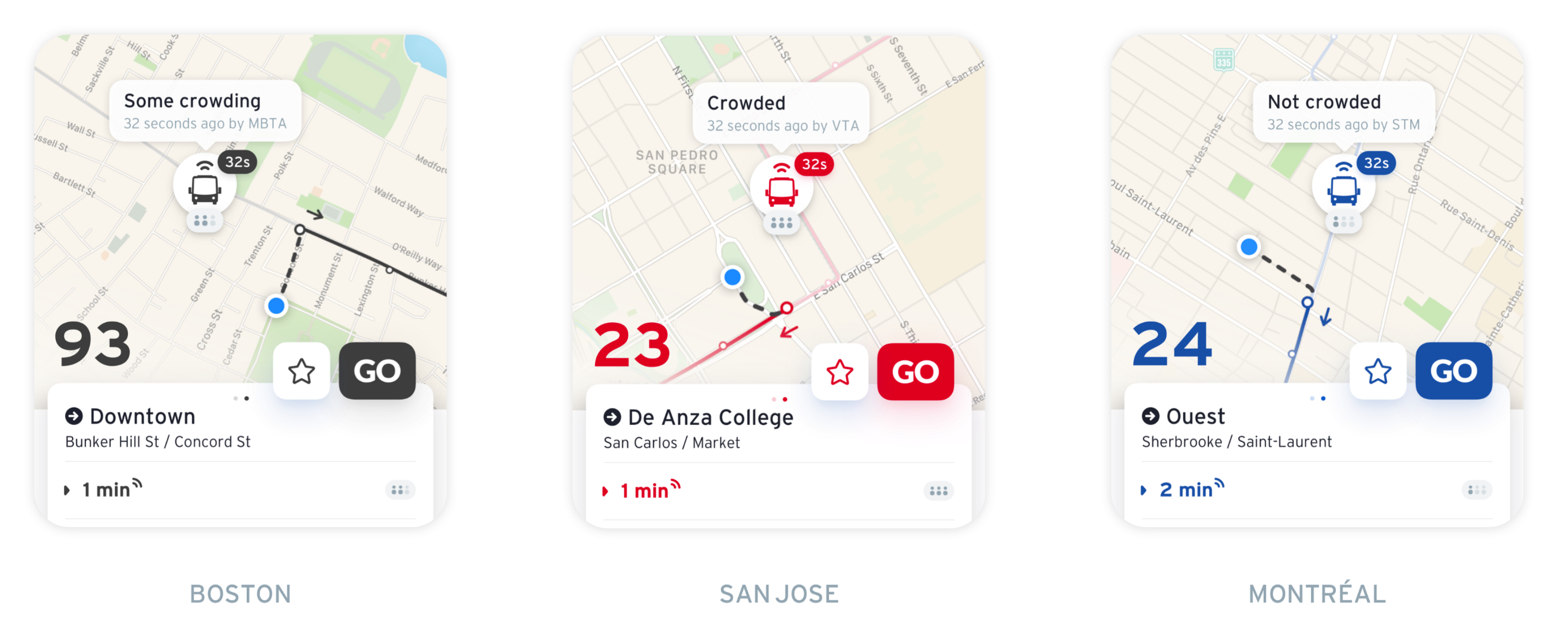
The Federal Transit Administration (FTA), where I work, has sought to address these challenges by promoting technology that can make public transportation easier and more seamless to use and more efficient to provide. Examples include:

*Smart-phone based transit navigation and vehicle arrival information:*

Apps like the one from the Transit company (shown below) use GPS technology to help passengers plan their travel using their smart phones, and show travelers the transit routes available, how long it will take to reach their destination, and when the next vehicle is scheduled to arrive. These apps sometimes also show non-transit options that are available, such as shared bikes and scooters.



Recently, apps such as transit have sought to provide travelers with real time information on how crowded a vehicle is:



However, in order for a rider to use this technology, transit agencies need to install automatic vehicle location technology on their vehicles and also make their route and vehicle information available in electronic format (often using the General Transit Feed Specification, or GTFS). In order to estimate vehicle crowding information, transit agencies need to install automatic passenger counters (APCs) on their vehicles.

*Electronic fare payment systems:*

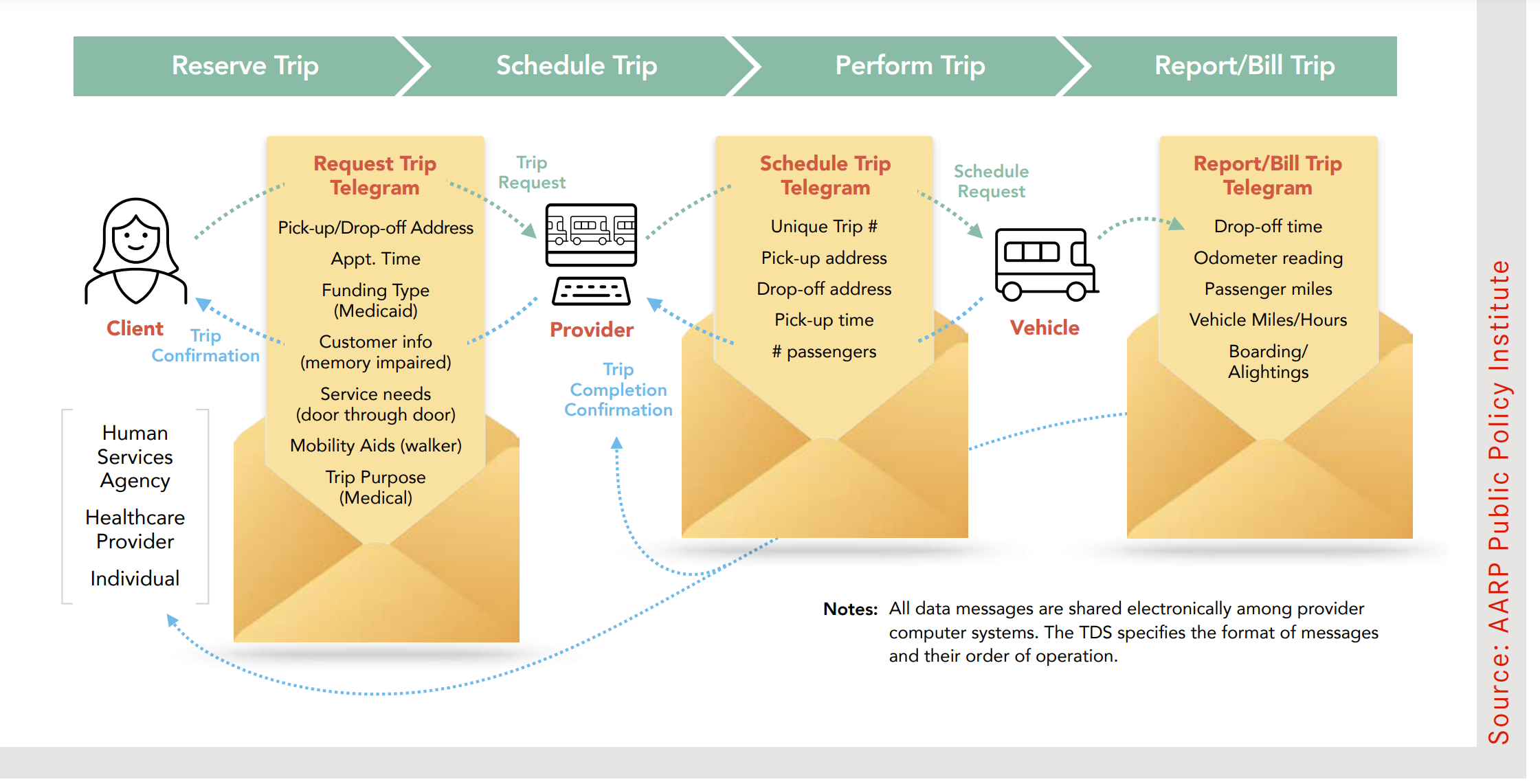
This technology is sometimes called “contactless fare payment”, “cashless fare payment,” or “smart trip” technology. It allows users to pay in advance for transit using a smart-phone app or electronic card. These technologies make it easier for people to use transit (no more waiting in line to buy a subway token or for the person in front of you to count out exact change) and make it more cost-effective for transit agencies to collect fares. They also make it easier for people using multiple modes of transportation (such as transit provided by different agencies or transit and bike sharing) since one type of fare payment can be used for all services.

During the COVID-19 pandemic, electronic fare payment has been a way to reduce the amount of surfaces a rider may touch and reduce the interaction between riders and bus drivers (which occurs when riders board the front of a bus to pay cash fare). In order for riders to use these technologies, agencies have to replace cash fare payment systems with electronic ones and sometimes have to partner with a software provider.

*Software that agencies use to coordinate public transit service:*

Most transit users never see this technology, but it can be very helpful for agencies to provide efficient service, especially for people who depend on door-to-door transit service. Under this service model, multiple human service providers use computer software and common data specifications to coordinate scheduling and dispatching to provide the most efficient trip to meet a client’s needs.

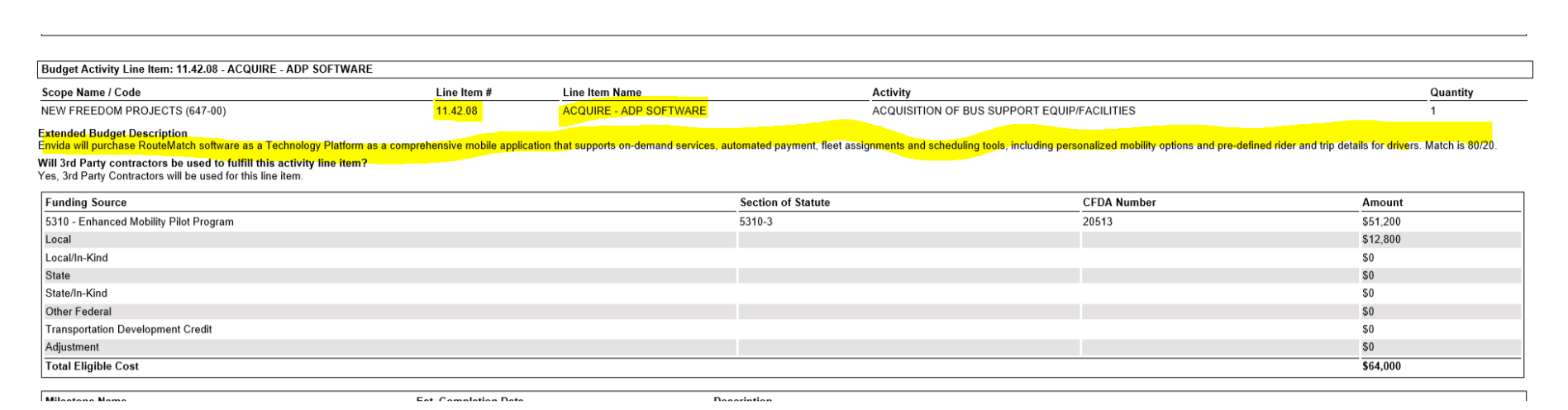
Image from: <https://www.aarp.org/content/dam/aarp/ppi/2020/12/modernizing-demand-responsive-transportation.doi.10.26419-2Fppi.00121.001.pdf>

These types of technologies are prevalent in transit systems in Asia and Europe, but not so common in the United States. FTA’s Office of Research, Demonstration, and Innovation (where I work) has been promoting the technologies for the past six years. We’ve funded demonstration projects and published the results, done research into advantages and barriers, and used conferences, webinars, and social media to champion these technologies. But we also know that adopting new technology can be expensive and there are many logistical and technological barriers that stakeholders must overcome to be successful.

We would like to know: ***have our efforts to promote innovative mobility technology been successful? How widely have these technologies been adopted by transit agencies?***

To help answer this question, this research uses data from grant applications submitted by transit agencies.

There are over 1,000 transit providers in the United States that receive direct financial assistance from the FTA. They range from very large agencies like the Metropolitan Transit Authority in New York, to much smaller providers in smaller towns that provide service with a half dozen buses. However, large or small, each agency needs to submit a grant application for Federal funding, and each application includes a budget that includes at least one budget Activity Line Item (ALI) and a narrative that describes the goods and services to be purchased. The narrative is called an Extended Budget Description (EBD). The ALI is a numerical code that is associated with quantitative data such as the amount of Federal funding requested. The EBD is a free text box. Here is an example of an ALI code and budget description (the ALI code and EBD are highlighted):



Unfortunately, FTA’s budget coding system have not caught up with the innovative mobility technology we have been sponsoring. Instead of providing specific numeric codes for travel navigation apps, electronic fare payment, and transit coordination software, our application management system provides general categories which are:

* Automated Data Processing (ADP) Hardware
* Automated Data Processing (ADP) Software
* Fare Collection Equipment (mobile)

Under these categories “software” could mean an agency is deploying an innovative smart-phone app to give travelers real time information about bus arrivals, or it could mean an agency is upgrading the Microsoft license it uses for its conventional office computers. Both projects are useful but only the first one identifies the kind of technology my office has been working to promote.

However, the EBD language contains more precise information on the purpose of the software, hardware, or fare equipment. Here is an example of an EBD for an innovative mobility project (the pertinent text is highlighted in yellow):

*The purpose of this project is to provide wireless infrastructure to HART facilities and Transfer Centers and Para-Transit Vans at HART. As HART moves forward with innovative solutions to technological enhancements for the benefit of our patrons, many solutions cannot be accommodated due to the lack of basic supporting infrastructure. A pivotal enhancement to support a myriad of technological advancements is the need to add wireless infrastructure (routers, supporting wiring, antennas, power supplies, etc.). This wireless infrastructure will be used to support a myriad of HART’s emerging technological enhancements including: new farebox technology which will allow the use of smartcard account based media and electronic media on personal electronic devices, future live streaming of CCTV, real-time location, ridership, and bus health information with the existing Orbital/AVL. An additional enhancement with this wireless infrastructure will be the ability for patrons to access the internet via their personal smart phone or other electronic devices with riding in HART vehicles.*

Here is an example of an EBD for more conventional technology:

*This Activity Line Item will fund the purchase of laptops and desktops used for remote/tele working for our Human Relations (HR), Accounting and Procurement staff that has been working from home during COVID-19. Our decision to have as many staff as possible to work remotely because on the onslaught of the COVID-19 pandemic. This was a way to ensure social distancing for our staff that did not need to be on site to do their work. We do not expect a return to the office for quite a while even after some restrictions have been lifted, in order to mitigate the spread of COVID-19 among our staff and for social distancing purposes that will keep our staff as safe as possible.*

And sometimes, the EBD doesn’t provide much information at all. Here is an example:

*ARCA to purchase 9 Smart Travel devices - $25,378.*

The problem is that we have thousands of EBD narratives, too many for one person to read in a short amount of time. Which leads to this paper’s central question:

***Can natural language processing and machine learning algorithms accurately classify innovative mobility technology vs. conventional technology?***

**Data Collection and Curation**

I downloaded data to a .csv file from FTA’s Transit Award Management System (TrAMS). I queried TrAMS for all grants made from Fiscal Year 2018 through Fiscal Year 2021 (as of March 31) that included one or more of the three budget ALI codes (ADP Hardware, ADP Software, Mobile Fare Collection Equipment).The resulting data consisted of 1,130 records including the EBD and additional information such as the grant program funding the purchase, the Federal funds and local matching funds, the name of the recipient, and the award date. Collectively the data represent 175 awards to 129 recipients for $93.7 million in Federal funding. (This dataset does not include information on transit agencies that adopted technology without Federal funds).

I generated a random sample of 200 records from the 1,130 using a random number generator and manually read each record, labeling the record “innovative” (1) if the record seemed to be for an innovative mobility and “conventional” (0) if the record seemed to be for traditional hardware, software, or fare payment.

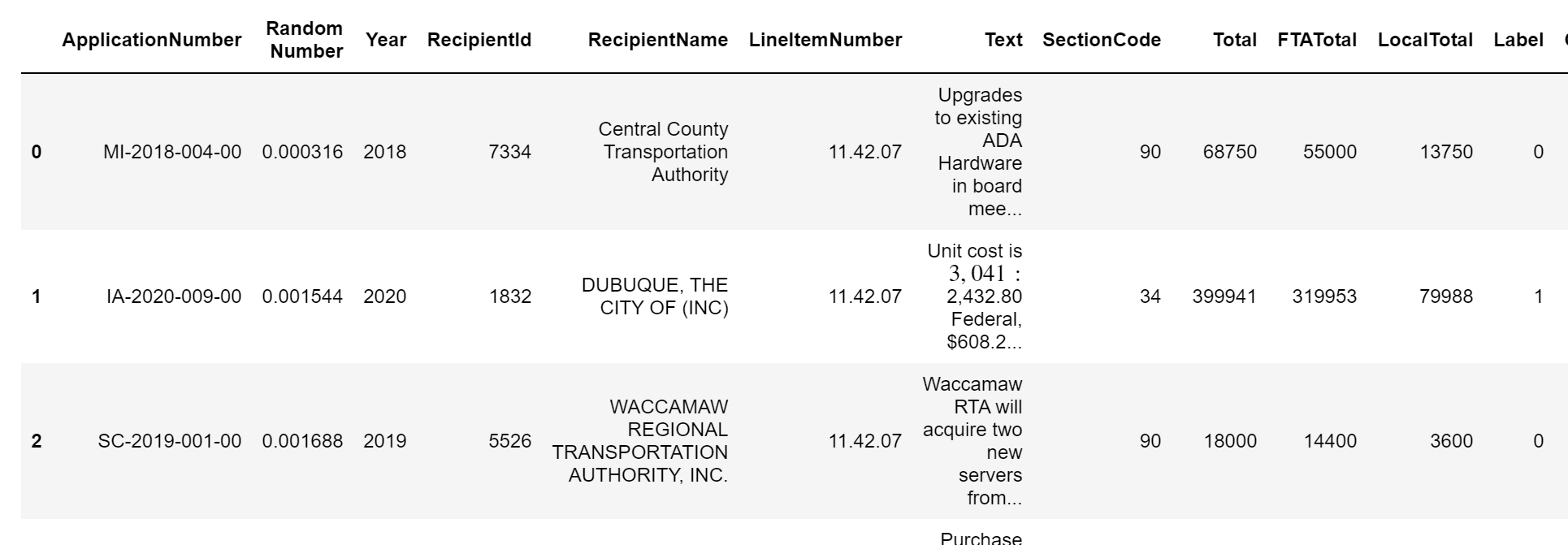
Here are some key words that triggered labeling an EBD as innovative or conventional technology:

|  |  |
| --- | --- |
| **Extended Budget Description Keyword Categorization** | |
| **"Innovative Technology"** | **"Conventional Technology"** |
| Dispatching software | Desktop computers |
| Mobile data terminals | Laptops |
| Automated Vehicle location | Tablets |
| Computer Assisted Dispatch | Hard drives |
| General Transit Feed Specifications | Payroll management systems |
| Real Time Information | Human Resource management software |
| Bus Location | Microsoft office products |
| Dynamic routing | Software licenses |
| Ridership projection software | Fare box machines on buses |
| Electronic fare payment | Fare box machines in stations |
| Cashless fare payment | Server |
| Touchless fare payment | Cyber Security |
| Smart Card system | Printer |
| RouteMatch software |  |
| Wireless infrastructure |  |
| Smart phone app |  |

If an EBD did not contain enough information to classify a technology, I classified it as “Conventional” (0).

**Preliminary Analysis and Data Cleaning**

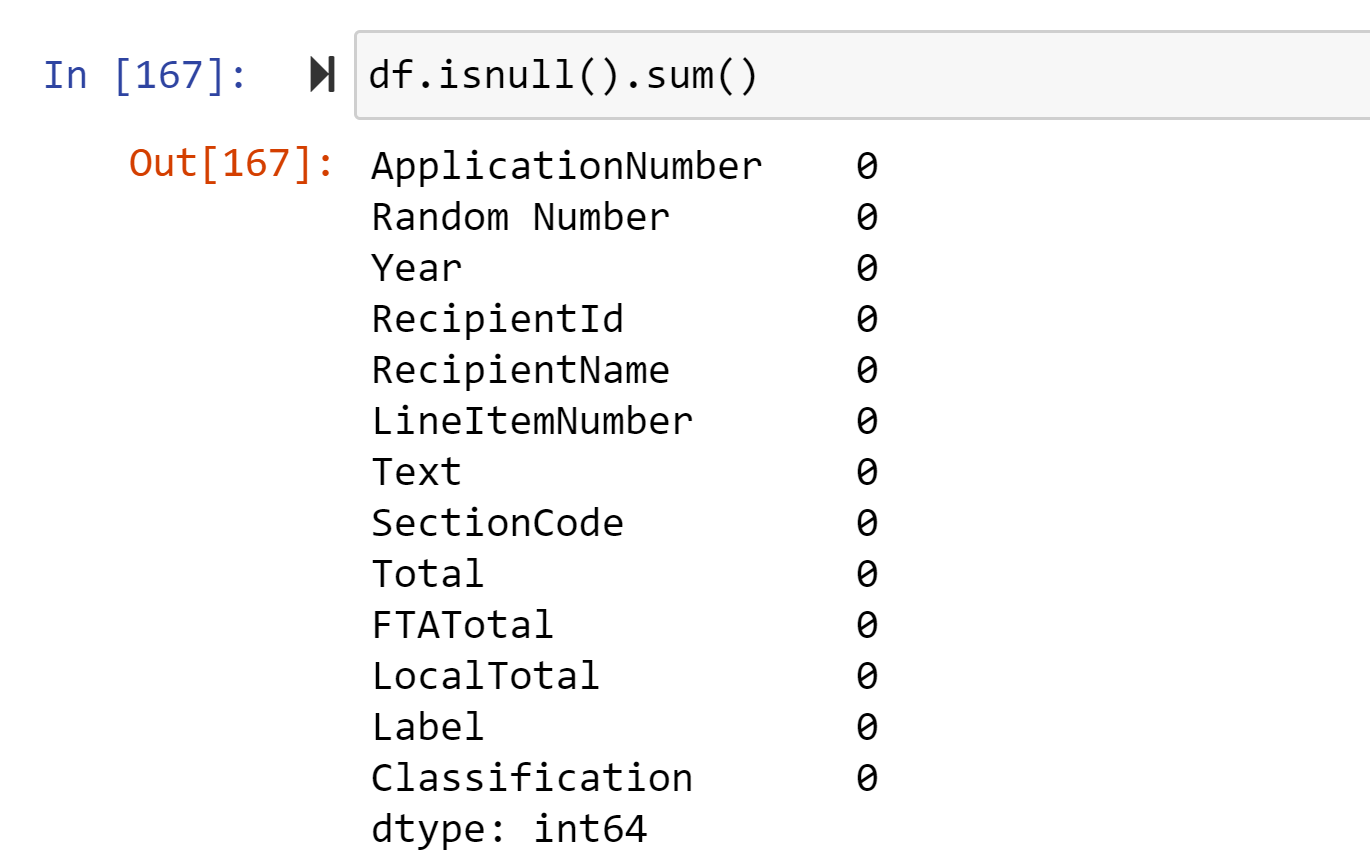
The data consists of 200 rows and 13 columns. Here is a snapshot of the first several rows of data:



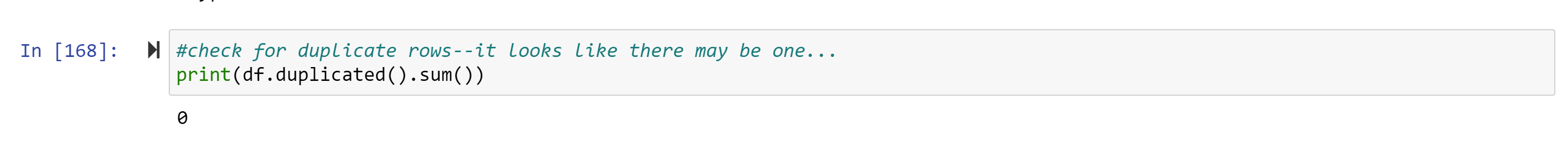
The data elements are:

* Application number: A unique ID assigned to each application for FTA grant funding
* Year: The fiscal year the application was awarded
* Recipient ID: A unique ID for each grant recipient
* Recipient Name: The Name of the Recipient
* Line Item Number: The budget ALI code that corresponds to either ADP Software, ADP Hardware, or Mobile Fare Collection
* Text: The EBD Text
* Section Code: A two digit number for the Federal program funding the award.
* Total: The combined total of Federal and non-Federal funds for the line item.
* FTA Total: The total amount of Federal funds for the line item
* Local Total: The total amount of non-Federal funds for the line item
* Label: Whether the line item represents conventional technology (0) or Innovative Technology (1)

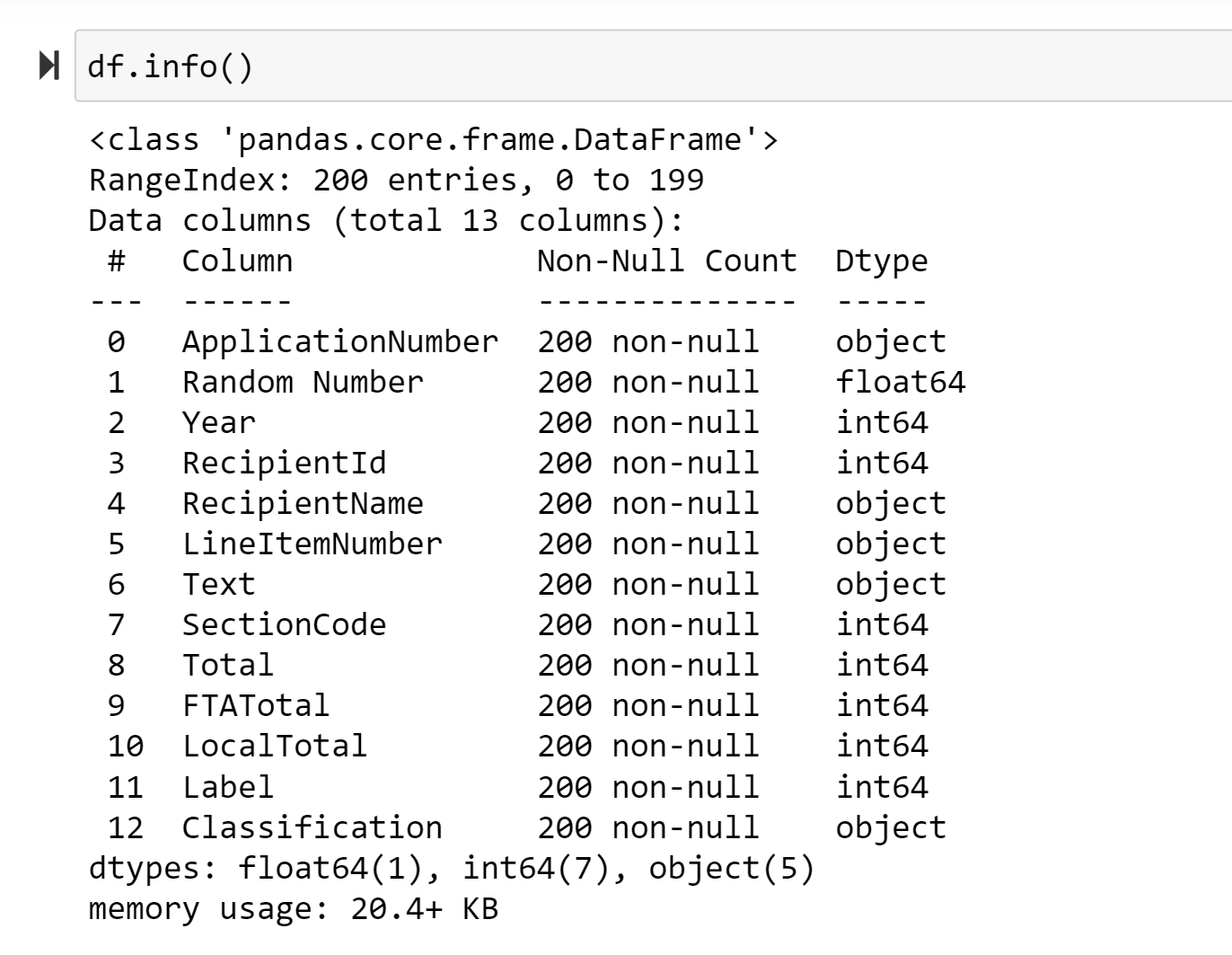
I checked the data to see if there was any missing information or duplicate information and there were no duplicates:



I also checked to see if there were duplicate rows and there were no duplicates:



Finally, I looked up the data types to see if they needed transformation. With the exception of the text category (which will be transformed later) all of the features were in formats that I could work with:



**Exploratory Data Analysis**

Before diving into the natural language processing, I explored the extent to which innovative technology was present in the labeled data, the amount of money spent on innovative vs. conventional technology, and other characteristics of the data. The tables below summarize my EDA code, which is available in the Jupyter Notebook.

The sample data was split roughly 60/40 between conventional and innovative technologies:

|  |  |  |
| --- | --- | --- |
| **Technology Type** | **Number** | **Percent of Total** |
| Conventional | 119 | 59.50% |
| Innovative | 81 | 40.00% |
| Total | 200 | 100% |

The mean Federal funding amount for conventional technologies is much higher than innovative technologies, but this is due to an outlier. The median funding amounts are similar:

|  |  |  |  |
| --- | --- | --- | --- |
| **Technology Type** | Mean Federal Funding | Median Federal Funding | Standard Deviation |
| Conventional | $586,797 | $60,305 | 4.57E+06 |
| Innovative | $295,289 | $57,568 | 6.83E+05 |

The number of recipients adopting innovative technologies and conventional technologies over the four year period is roughly the same, with slightly more recipients adopting conventional technologies. 42 recipients adopted both innovative and conventional technologies between 2016 and 2021.

|  |  |
| --- | --- |
| **Category** | **Total** |
| Recipients Adopting Conventional Technology | 80 |
| Recipients Adopting Innovative Technology | 72 |
| Recipients Adopting Both Types of Technology | 40 |

When it comes to technology adoption over time, the number of conventional technologies stayed relatively constant but the number of innovative technologies increased from 2018-2020 (since 2021 was only halfway over at the time of this analysis, it’s not surprising to see a drop-off):

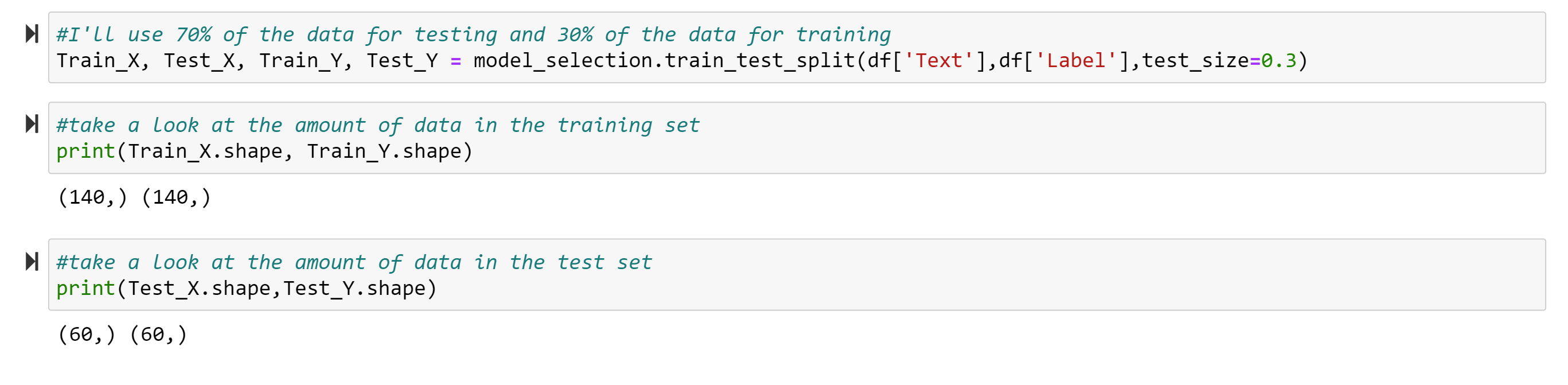
|  |  |  |
| --- | --- | --- |
| Year | Conventional | Innovative |
| 2018 | 40 | 4 |
| 2019 | 40 | 23 |
| 2020 | 39 | 41 |
| 2021 | 1 | 1 |
| Grand Total | 119 | 81 |

The proportion of conventional and innovative technologies funded by different programs is about the same, with the Section 90 program (known as the Urbanized Area Formula Program) funding most of both types of technology:

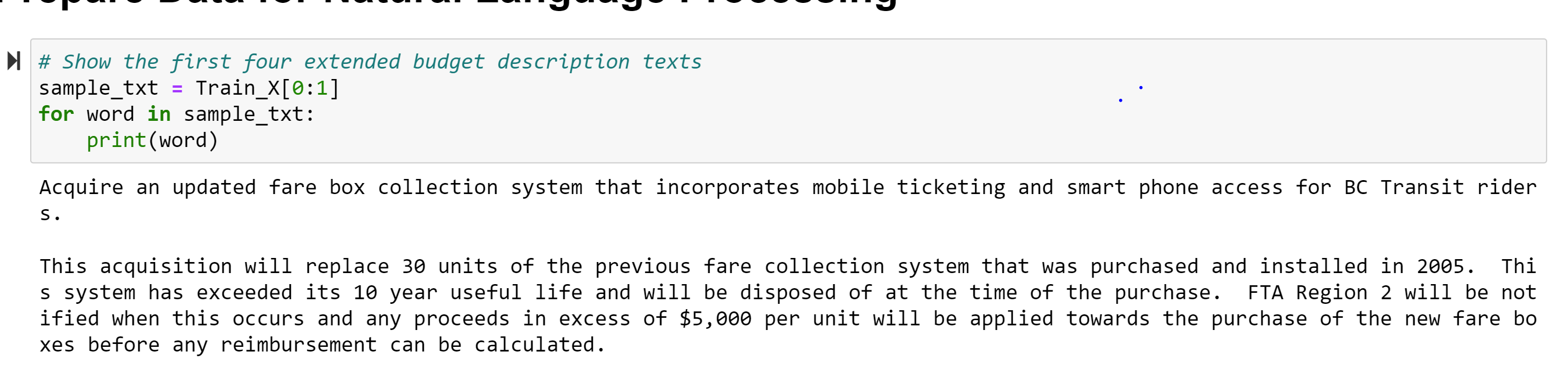
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Program Name** | **Conventional Technology** | **Share of total** | **Innovative Technology** | **Share of Total** |
| **90** | **79** | **66%** | **43** | **53%** |
| **34** | **13** | **11%** | **16** | **20%** |
| **18** | **12** | **10%** | **3** | **4%** |
| **16** | **7** | **6%** | **13** | **16%** |
| **54** | **5** | **4%** | **1** | **1%** |
| **95** | **2** | **2%** | **3** | **4%** |
| **65** | **1** | **1%** | **1** | **1%** |
| **79** | **0** | **0%** | **1** | **1%** |
| **Total** | **119** | **100%** | **81** | **100%** |

**Preparing the Data for Natural Language Processing**

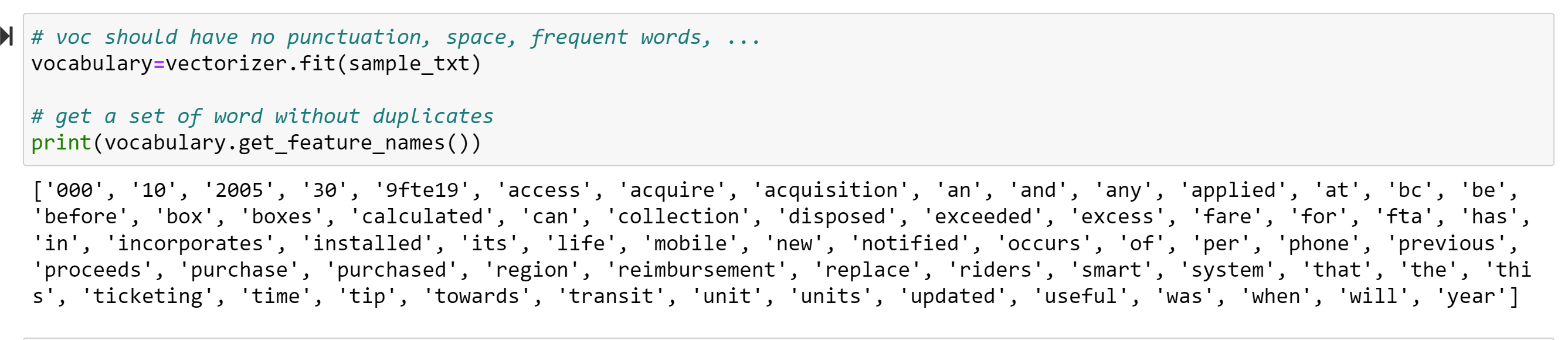
I split the data into training and test sets, with 70% of the data for training and 30% set aside for testing.



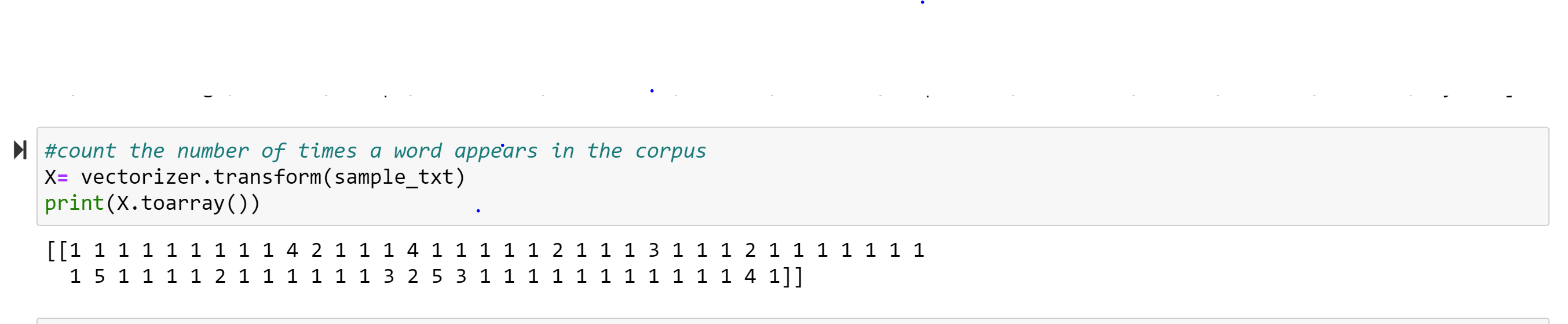
I then called the Vectorizor function to convert the text into numbers. Here is an example of text prior to vectorization:



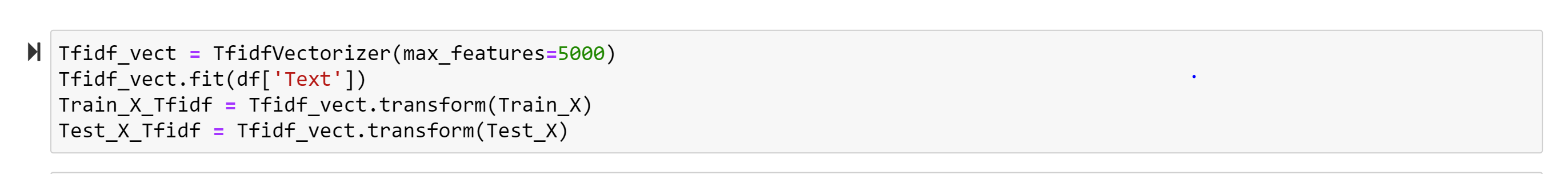
And here is the same text after Vectorization:



Finally, here are the number of times each word appears in the document:

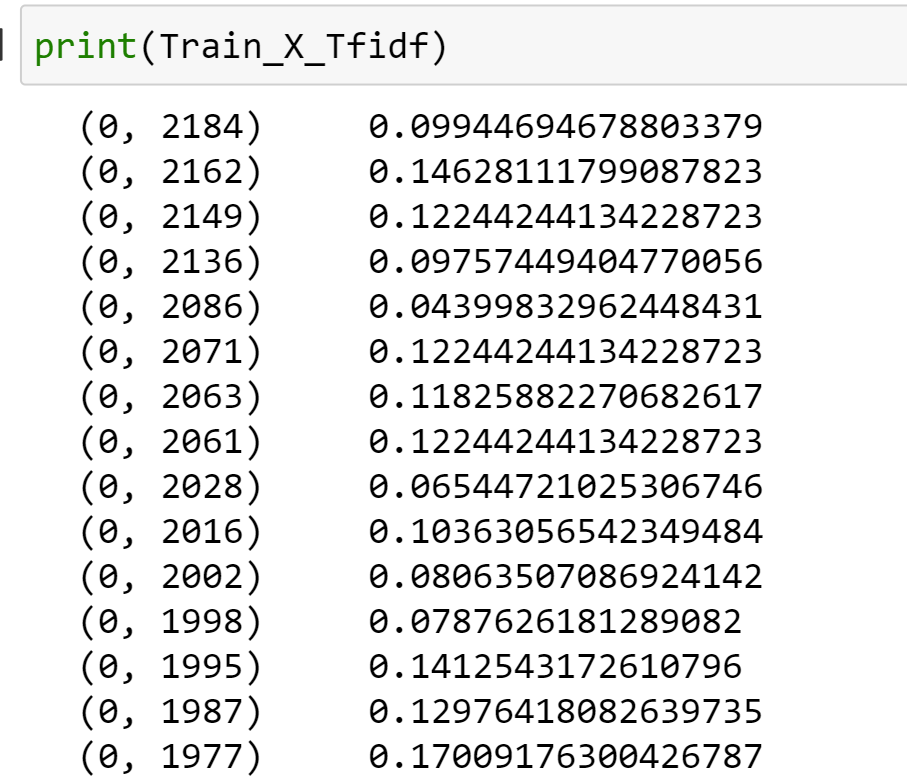


Then I initiated a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer to assign scores to each word. The TF-IDF code will assign word frequency scores that try to highlight words that are more interesting, e.g. frequent in a document but not across documents.



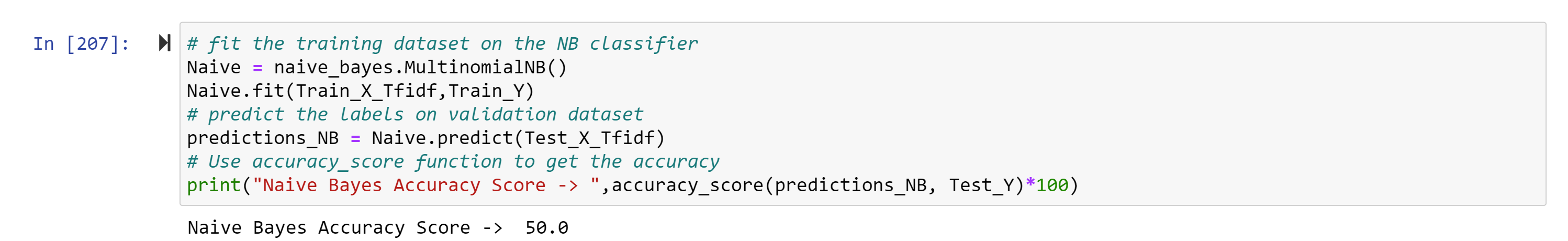
This will help TF-IDF build a vocabulary of words which it has learned from the corpus data and it will assign a unique integer number to each of these words. There will be maximum of 5000 unique words/features since parameter max\_features=5000.

Below is a partial printout of the vectorized data. The first number in the parentheses represents the row number, the second number represents the unique Integer number of each word in the first row, and the number outside the parentheses is the score calculated by TF-IDF Vectorizer.

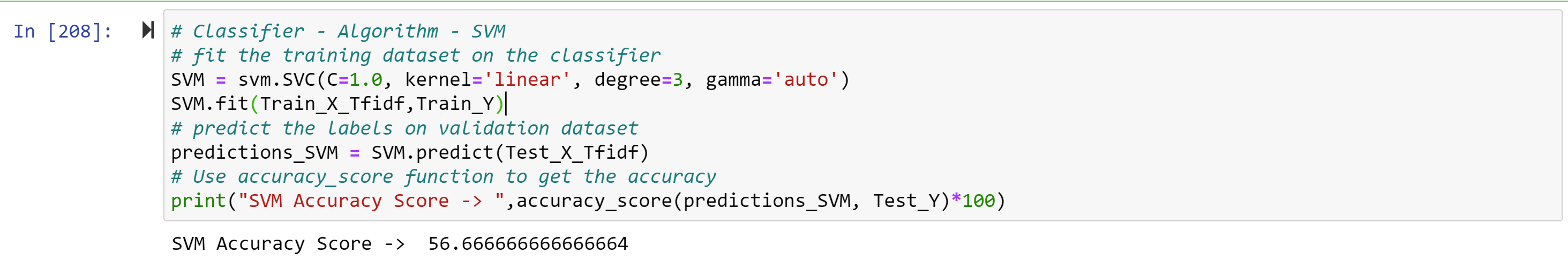


**Binary Classification using Scikit Learn Classifiers**

Now that I’ve prepared the data for NLP classification, I’ll test out the training set on some common scki-kit learn classifiers. The first one I try is a Naïve Bayes classifier, but it produces an accuracy of only 50%, no better than classifying the text as “innovative” or “conventional” based on a coin flip:



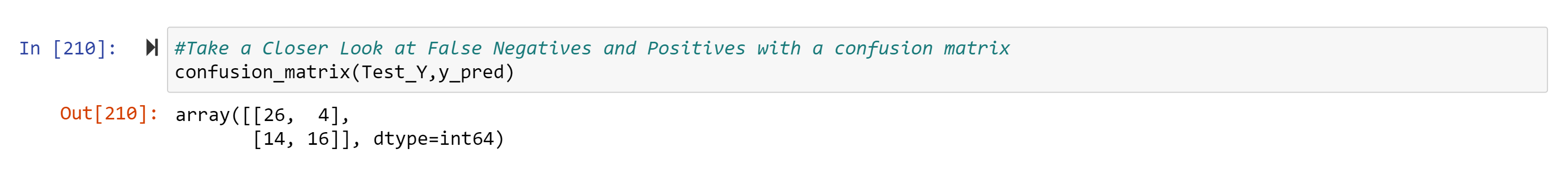
I also try out a Support Vector Machine (SVM) classifier, which gives me slightly better results, but not that much better:



Finally, I try out a PassiveAgressive classifier, which gives me results of 70%. Better, but not great.



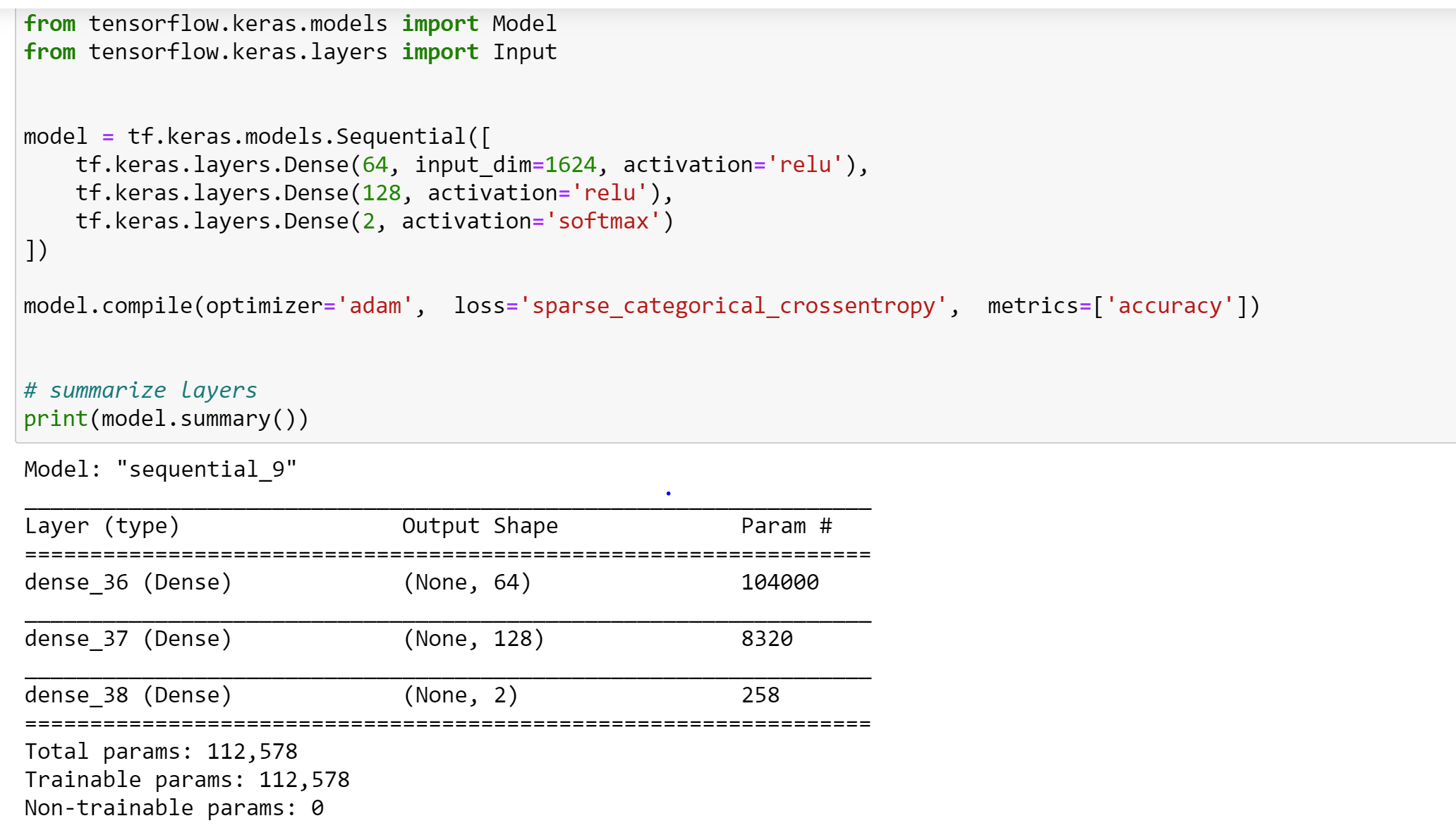
I’m curious about whether the PassiveAggressive classifier is erring on the side of producing false positives or false negatives, so I run a confusion matrix on the test set:



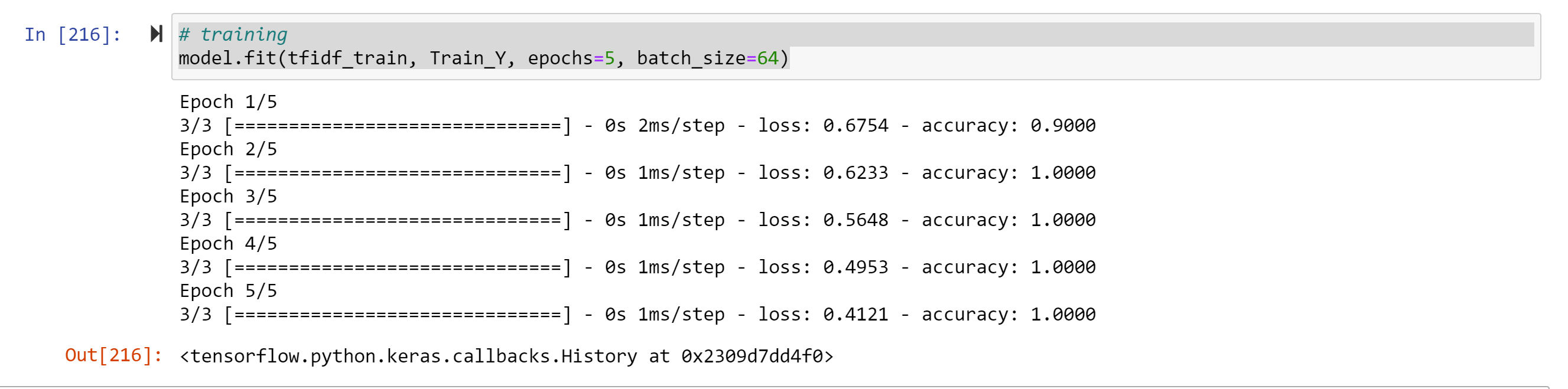
It looks like the false negatives (14) outweigh the false positives (4) meaning the classifier is more frequently classifying innovative technology as conventional than the other way around. Since we want to be conservative and err on the side of undercounting innovative technology, this could be OK. But with only a 70% accuracy rate, I might just want to read and label the remaining 900 records manually.

**Binary Classification using Deep Neural Networks**

I next try to classify the training and test data using a simple Deep Neural Network with one hidden layer:



However, once I run the DNN, it’s pretty clear that the model is overfitting the training data. I get an accuracy of 100% after the first epoch:

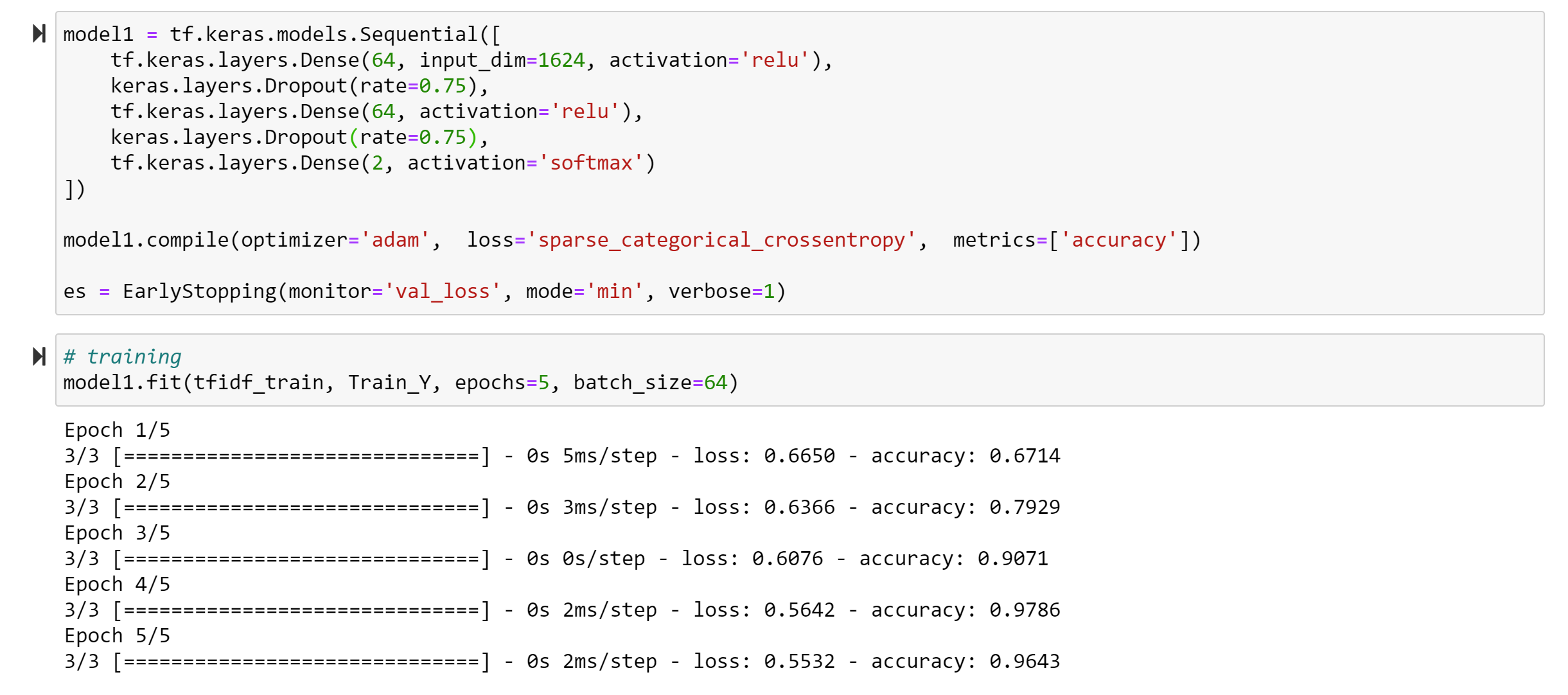


**Adjusting for Overfitting**

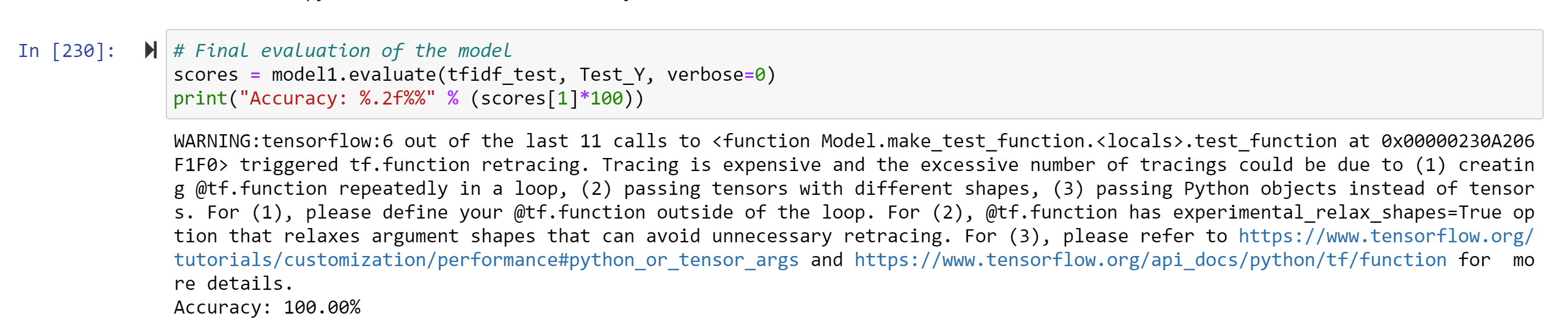
To adjust for overfitting I try a combination of tactics:

* Adding a dropout
* Adding early stopping.

These techniques get me to an accuracy of 96%



However, when I run the DNN on the test set, I get a 100% accuracy rate and a strange warning message:



I suspect I may not have run the model correctly and the DNN is still overfitting the data, though the results seem more promising then the sci-kit learn models ran previously.

**Conclusion**

My Exploratory Data Analysis indicates that transit agency adoption of innovative mobility technologies is relatively widespread, with an almost equal number of recipients adopting innovative vs. conventional technologies since 2018 and with innovative technologies growing over time. Assuming my sample of 200 records is representative, widespread adoption should be present in the population as well. Natural Language Processing algorithms proved to be a limited tool to predict whether a technology was conventional or innovative. It’s possible my sample size was too small or the amount of text or complexity of text in the EBD was not rich enough to allow for accurate binary classification.

The obvious next step would be to create separate budget codes for innovative mobility technologies. Absent that step, my agency could give guidance to grant applicants to include more details in the EBD to allow for a better understanding of the types of technologies being implemented by public transportation.