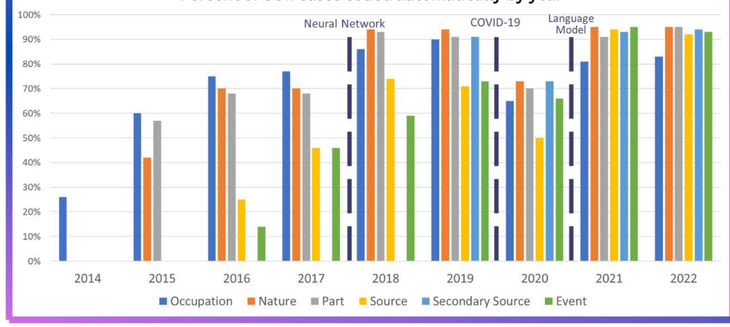
Autocoding of Survey Data at the Bureau of Labor Statistics

Daniel Todd, Sertan Akinci, Mohamed Moulaye
OCWC/CRPDG, OEWS/OAC

What is the SOII Autocoder?

The Office of Compensation and Working Conditions (OCWC) receives over 200k free-text injury and illness case narratives for the Survey of Injuries and Illnesses (SOII) each year. To publish statistics on these cases, each must be assigned Occupational Injury and Illness Classification System (OIICS) codes as well as Standard Occupational Classification (SOC) codes. This was historically done by humans, which can be time consuming and prone to error. With the introduction of machine learning models, the SOII can code a large majority of its cases automatically, saving time and increasing consistency.

Percent of SOII cases coded automatically by year



Example Narrative

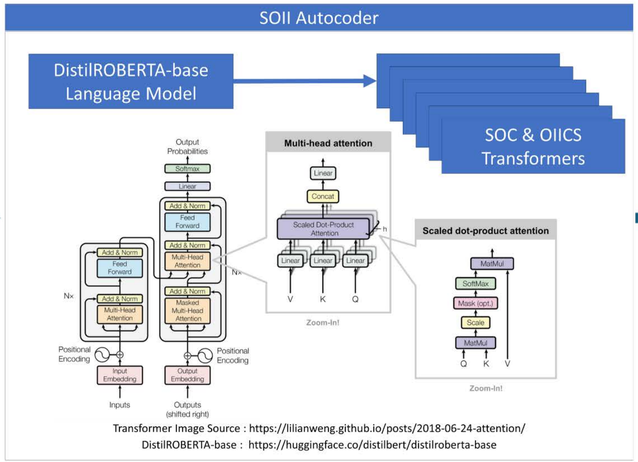
Job title: Sanitation worker

What was the employee doing just before the incident?
Mopping floor in gym

What happened?
Slipped on wet floor and fell

What part of the body was affected?
Fractured right arm

What object directly harmed the employee?
Wet floor



Transformer Image Source : https://lilianweng.github.io/posts/2018-06-24-attention/
DistilROBERTA-base : https://huggingface.co/distilbert/distilroberta-base

Codes Assigned

SOC: 37-2011 (Janitor)
SOC Prob: 64%
OIICS-Nature: 124 (Fracture)
Nature Prob: 95%
OIICS-Part: 420 (Arm)
Part Prob: 98%
OIICS-Event: 4312 (Fall, slipping)
Event Prob: 52%
OIICS-Source: 6624
(Other constructed surface)
Source Prob: 57%
OIICS-Secondary Source: 4152 (Water)
Secondary Source Prob: 88%

Human Review

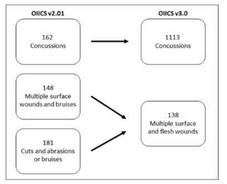
All codes assigned by the Autocoder are reviewed by humans prior to being used in the SOII.

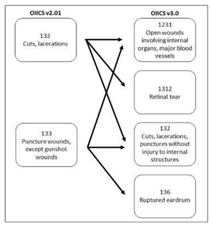
Probability Thresholds

Probability thresholds are used to determine when the Autocoder should not be trusted, resulting in only codes which exceed those thresholds being shown to human reviewers, and those that dont meet thresholds are assigned manually.

OIICS v3.02

Starting in Survey Year 2023, the OIICS coding system was updated from version 2.01 to version 3.02. To allow the Autocoder to predict these codes, all training data was crosswalked from v2 to v3 with two common mappings shown below - one-to-one (left) and one-to-many (right).





What is the OEWS Autocoder?

The Occupational Employment and Wage Statistics (OEWS) program samples up to 200,000 establishments every six months and the state workforce agencies and BLS regional offices process millions of occupational descriptions to assign them with a Standard Occupational Classification (SOC) code so they be processed in our systems. In order to assist with this workload, the OEWS national office created the Autocoder system, a machine learning model that predicts SOC codes for a given occupation description.

Autocoded Titles

History of Models

1) SGDClassifier : This is our first and current machine learning model that uses logistic regression to train and stochastic gradient descent to optimize. The data is first vectorized then fed into the model for training. There are around 8 million records in the training data. We use variables like job title, company identifier and industry code to predict occupation codes.

2) Convolution Neural Network: Our second and soon to be the next model is CNN. As shown on the right, job titles are vectorized and then are used to create embedding matrixes before being fed into the convolution layer.

3) RRN and Transformers: We are also working on a recurrent neural network (RNN) and a transformer for our next model update.

Accuracy Monitoring

Only the SOC codes above a certain confidence level are sent to the states. Periodically the Autocoder is trained with more data and the threshold is updated based on the Autocoder' s accuracy performance. This performance is measured by expert state coders to achieve a gold standard for this process. "Gold-Coded" dataset used to separately test the accuracy of the Autocoder and determine thresholds. The Gold-Code dataset contains job titles that are expertly coded and agreed upon by two state coders without assistance from the autocoder. This process brings an impartial review of overall coding accuracy of both the autocoder and human coders. Current Gold-Code file contains around 76,000 titles.

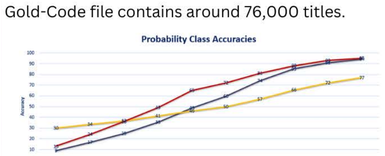
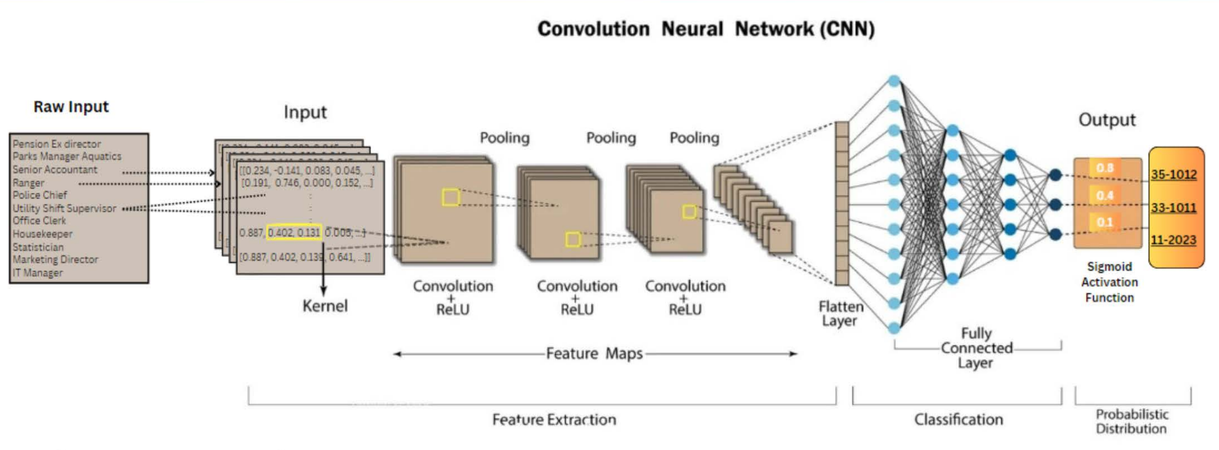


Table 10

		When First Code Correct		When Second Code Correct			
Probability Class	Code Choice	Total # of Titles	Code Choice Distribution	Total # of Titles	Code Choice Distribution	Type-1 False Positive	Type-2 False Negative
9 (.8 - .9)	Neither	2,262	19.8	64	23		
9 (.8 - .9)	First Autocode	8,862	77.7	91	32.7		
9 (.8 - .9)	Second Autocode	281	2.5	123	44.2		
Total		11,405		278		22.3	55.8
10 (.9 - 1)	Neither	992	17.9	2	5.9		
10 (.9 - 1)	First Autocode	4,484	80.8	19	55.9		
10 (.9 - 1)	Second Autocode	75	1.4	13	38.2		
Total		5,551		34		19.2	61.8

Convolution Neural Network (CNN)



CONTACT INFORMATION

DANIEL TODD : TODD.DANIEL@BLS.GOV
MOHAMED MOULAYE : MOULAYE.MOHAMED@BLS.GOV
SERTAN AKINCI : AKINCI.SERTAN@BLS.GOV

RESOURCES

DOL AI USE CASE LIBRARY : HTTPS://WWW.DOL.GOV/AGENCIES/OASAM/CENTERS-OFFICES/OCIO/AI-INVENTORY
OCWC : HTTPS://WWW.BLS.GOV/IIF/
OEWS : HTTPS://WWW.BLS.GOV/OES/