sarahlintang@gmail.com

Task Definition

In Leadbook there are any millions of contacts and hundreds of thousands new contacts added daily. One of the important tasks is to classify the department(s) from job title of a contact. In general a rule based algorithm with bag of words is used to classify the job title to a department however with the increasing variety of job titles and new job titles created every year, the complexity of words and classification increases. How can we solve this problem using machine learning. Y

Approach

From the task definition we conclude that this case can be solved using unsupervised learning and supervised learning. We utilize BERT embedding for this case, because BERT offers an advantage over models like Word2Vec. BERT produces word representations that are dynamically informed by the words around them

1. Unsupervised

   Since the dataset given by LeadBook have no label, the easy way is using unsupervised learning. Here we apply BERT to classify text by word vector similarity.

   - Clean data and embed it into the vector space
   - Create a topic cluster for each category and embed it into the vector space
   - Calculate similarities between every text vector and the topic clusters, then assign it to the closest cluster.

   In order to evaluate the model, we used job title dataset gathered from O*NET then we run the usual evaluation metrics (Accuracy, Precision, Recall).

2. Supervised

   To compare with unsupervised, we train the supervised learning with deep learning using BERT. Basically we only simple classifier layer (Feed Forward Neural Network + Softmax) to give class probabilities of the job title.

   We used Google Colabs to run this BERT classification. Because we need GPU for faster training.

   Configuration:

   - Dataset:

     To train the model we used gathered dataset from O*NET. But we did not use all the dataset due the imbalance dataset. Therefore we adjust to 650 data every class.
   - Optimezer: AdamW

     We chose Adam optimizer because some resources said Adam works well in practice and compares favorably to other stochastic optimization methods. While AdamW is an improved version of Adam.
   - Batch-size: 16

     Because we used GPU and the dataset not really big, we chose 16 for batch size.

- Epoch : 10
  We set 10 epoch because after 6 epoch there is no accuracy improvement.

Data

Labeled dataset was needed to evaluate the model of unsupervised and supervised. Some dataset was gathered from O*NET and got 53.594 data contains job titles and SOC code. We cluster the data based on an alternative aggregation suggested by SOC-2018 with an intermediate classification level as shown table 1.

Table 1. Clustering Dataset based on SOC-2018

Intermediate Aggregation	Major Groups Included	Intermediate Aggregation Title
1	11 – 13	Management, Business, and Financial Occupations
2	15 – 19	Computer, Engineering, and Science Occupations
3	21 – 27	Education, Legal, Community Service, Arts, and Media Occupations
4	29	Healthcare Practitioners and Technical Occupations
5	31 – 39	Service Occupations
6	41	Sales and Related Occupations
7	43	Office and Administrative Support Occupations
8	45	Farming, fishing, and Forestry Occupations
9	47	Construction and Extraction Occupations
10	49	Installation, Maintenance, and Repair Occupations
11	51	Production Occupations
12	53	Transportation and Material Moving occupations
13	55	Military Specific Occupations

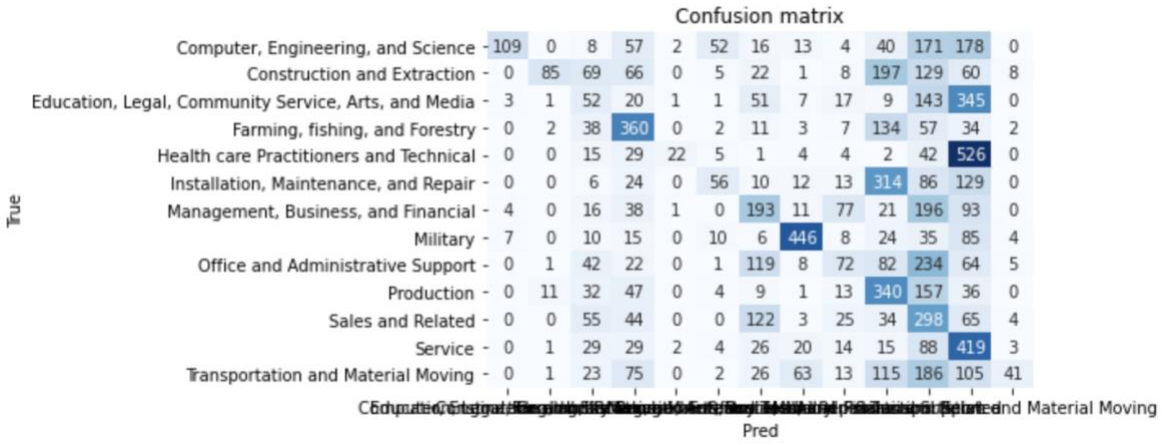
Source: (Bureau of Labor Statistics 2018)

Result

1. Unsupervised

```
F-1 Score: 0.27
 Accuracy: 0.3
 Auc: 0.5
 Detail:
```

	precision	recall	f1-score	support
Computer, Engineering, and Science	0.89	0.17	0.28	650
Construction and Extraction	0.83	0.13	0.23	650
Education, Legal, Community Service, Arts, and Media	0.13	0.08	0.10	650
Farming, fishing, and Forestry	0.44	0.55	0.49	650
Health care Practitioners and Technical	0.79	0.03	0.06	650
Installation, Maintenance, and Repair	0.39	0.09	0.14	650
Management, Business, and Financial	0.32	0.30	0.31	650
Military	0.75	0.69	0.72	650
Office and Administrative Support	0.26	0.11	0.16	650
Production	0.26	0.52	0.34	650
Sales and Related	0.16	0.46	0.24	650
Service	0.20	0.64	0.30	650
Transportation and Material Moving	0.61	0.06	0.11	650
accuracy			0.30	8450
macro avg	0.46	0.30	0.27	8450
weighted avg	0.46	0.30	0.27	8450



Test the model on dataset given by LeadBook:

	job title	dept prediction
0	art	Management, Business, and Financial
1	interior architect	Office and Administrative Support
2	supervisor call centre	Management, Business, and Financial
3	tv host	Education, Legal, Community Service, Arts, and...
4	senior fuel trader	Sales and Related
5	director global operation program management o...	Sales and Related
6	operation executive	Sales and Related
7	executive assistant manager sale marketing ser...	Management, Business, and Financial
8	professional tennis coach	Management, Business, and Financial
9	business	Sales and Related
10	senior credit analyst	Sales and Related
11	product	Sales and Related
12	supervising senior engineer	Service
13	marketing operation manager	Management, Business, and Financial
14	diploma mass communication student	Office and Administrative Support
15	freelance editor writer	Management, Business, and Financial
16	supply oceania	Production
17	lead commissioning engineer	Service
18	casino audit senior officer	Service
19	sl manager	Farming, fishing, and Forestry

2. Supervised

Here the F-1 Score of O*NET dataset.

sarahlintang@gmail.com

```
Epoch 6
Training loss: 0.2558776523279055
Validation loss: 0.5220054022269324
F1 Score (Weighted): 0.8547726471647185
```

Validation Test

```
Class: 0
Accuracy: 81/97

Class: 1
Accuracy: 86/98

Class: 2
Accuracy: 84/97

Class: 3
Accuracy: 91/98

Class: 4
Accuracy: 83/98

Class: 5
Accuracy: 74/97

Class: 6
Accuracy: 81/98

Class: 7
Accuracy: 82/98

Class: 8
Accuracy: 83/97

Class: 9
Accuracy: 80/97

Class: 10
Accuracy: 89/98

Class: 11
Accuracy: 78/97

Class: 12
Accuracy: 91/98
```

Test on dataset from LeadBook

	job title	job clean	dept prediction
0	director, operations and strategic projects	director operation strategic project	Management, Business, and Financial
1	head supply chain management, group emerging m...	head supply chain management group emerging ma...	Sales and Related
2	systems/software engineer	engineer	Computer, Engineering, and Science
3	apac director, industrial air filtration	director industrial air	Transportation and Material Moving
4	infopreneur and internet marketer	internet	Computer, Engineering, and Science
5	assistant vice president team manager	assistant vice president team manager	Management, Business, and Financial
6	bookroom manager	manager	Management, Business, and Financial
7	grossly under employed	employed	Education, Legal, Community Service, Arts, and...
8	qa automation	automation	Installation, Maintenance, and Repair
9	graphic designer & large format printing operator	graphic designer large format printing operator	Production
10	director (tax - global mobility)	director tax global mobility	Management, Business, and Financial
11	assistant finance director	assistant finance director	Management, Business, and Financial
12	director, finance and corporate services	director finance corporate service	Management, Business, and Financial
13	guitar lessons	guitar lesson	Education, Legal, Community Service, Arts, and...
14	sports therapist	sport therapist	Health care Practitioners and Technical
15	information technology (regional) mgr	information technology regional	Computer, Engineering, and Science
16	senior supervisor, apac region plant engineering	senior supervisor region plant engineering	Production
17	director - commodities	director commodity	Management, Business, and Financial
18	senior volunteer	senior volunteer	Education, Legal, Community Service, Arts, and...
19	managing director - asia pacific	managing director asia pacific	Management, Business, and Financial

From the result above, we can see that the dept prediction was pretty good. for the future, we can improve the model using more dataset for training. Actually, we can train all the dataset collected from O*NET. Due to it really takes time, so for this assignment we just used 650 data for every class. But we are sure if all the dataset is used, the result will be better.

Conclusion

From the results of supervised and unsupervised we conclude that the supervised BERT classification gives much better result. So, the model from supervised can be used for future. The classification class was based on Standard Occupation Classification (SOC) so even with increasing variety of job titles and new job titles created every year, this classification still valid. Thus, we still can use this model.