PGM PROJECT - DELIVERABLE 3

Mohammad Nur AMIN **Jatin BABBAR** **Shirin BASHIRIAMID** **Madhubhani RANCHA GODAGE**

Orhan SOLAK

1 Evaluation

In this paper, we discuss about the evaluation details of all the models we implemented to address the problem of job classification.

1.1 Multinomial Naive Bayes

class	precision	recall	f1-score	support
0	0.71	0.75	0.73	225
1	0.92	0.50	0.65	225
2	0.77	0.86	0.81	225
3	0.60	0.62	0.61	225
4	0.76	0.81	0.78	225
5	0.71	0.77	0.74	225
6	0.76	0.72	0.74	225
7	0.78	0.65	0.71	225
8	0.95	0.79	0.86	225
9	0.87	0.80	0.83	225
10	0.81	0.74	0.77	225
11	0.53	0.68	0.59	225
12	0.78	0.91	0.84	225
13	0.49	0.92	0.64	225
14	0.69	0.90	0.78	225
15	0.63	0.48	0.55	225
16	0.87	0.82	0.84	225
17	0.84	0.38	0.52	225
18	0.87	0.73	0.79	225
19	0.73	0.87	0.80	225
20	0.77	0.80	0.79	225
21	0.52	0.40	0.46	225
22	0.71	0.87	0.78	225
23	0.88	0.70	0.78	225
24	0.61	0.69	0.65	225
25	0.74	0.84	0.78	225
26	0.86	0.71	0.78	225
27	0.78	0.75	0.76	225
accuracy			0.73	6300
macro avg	0.75	0.73	0.73	6300
weighted avg	0.75	0.73	0.73	6300

Table 1: Multinomial Naive Bayes results.

Basic Multinomial Naive Bayes, gives 0.75 *Precision* 0.73 *F1-score* as weighted average. It is not so much good result, but considering the simplicity and small execution time (0.8 seconds) its applicable and good method.

1.2 Latent Dirichlet Allocation

LDA is evaluated by estimating the probability of unseen document given training documents. Our model learn concepts based on categories and words correspond them to recognize the correct category of a new documents based on its words.

It learns the category distribution based on the given fix number of categories. Each description in dataset has the category distribution.

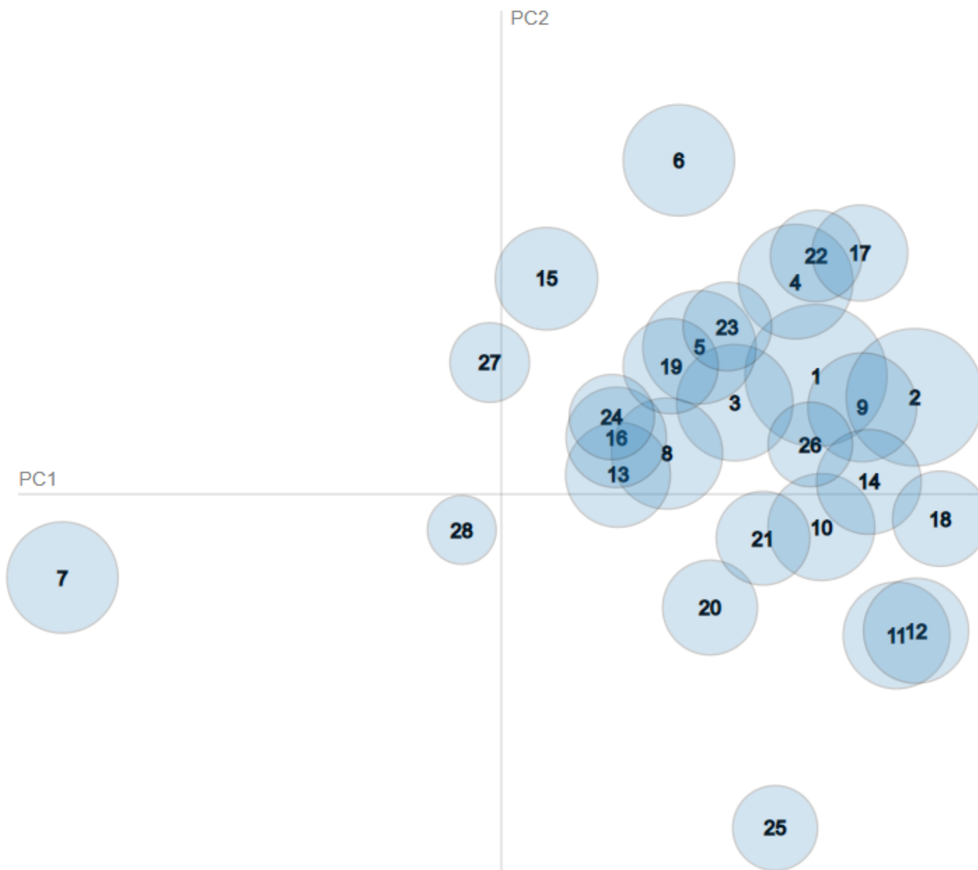

Figure 1: Intertopic distance map

We analyze our model in an interactive graphical representation (`https://ztccybwao5evrcebtdfuqa-on.drv.tw/PGM/ldavis_prepared_28.html#topic=0&lambda=1&term=`LDA visualization). By referring Intertopic distance map, we can see how topics are related to each other and by referring top 30 most salient terms graph, we can learn the most important or frequent words for each topic.

1.3 Gibbs Sampling

For the experimentation, we randomly kept around only 20% of the labels and the rest were hidden. We sampled for 20 iteration and from the Figure 2 we can visualize the original labels vs the predicted labels for reduced 210 samples for two components for better visualization.

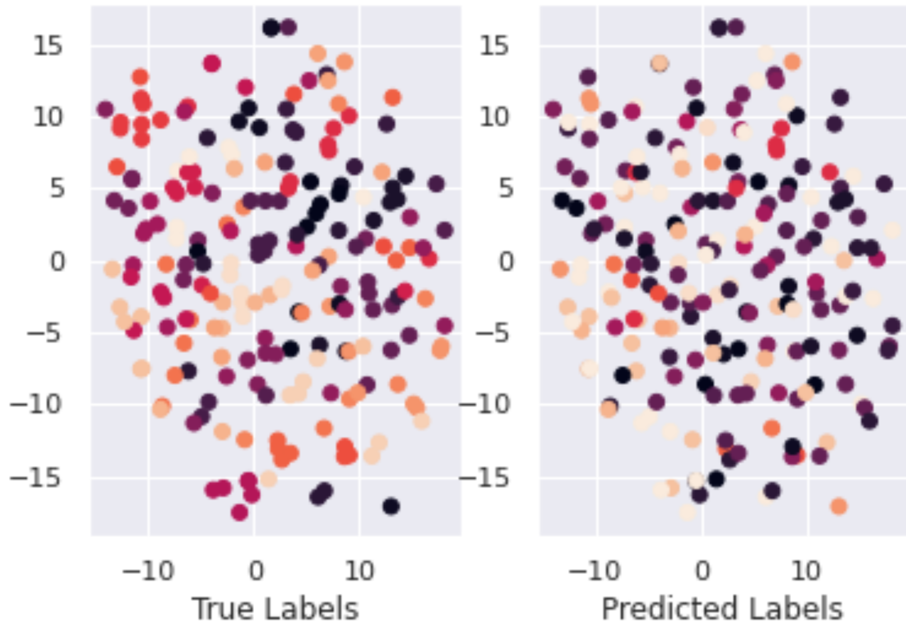

Figure 2: tSNE representation for document labels

We can visualize the confusion matrix for our predictions in Figure 3 and we can see that most of the entries are along the diagonal which is expected. This not not the best fit given the nature of the data.

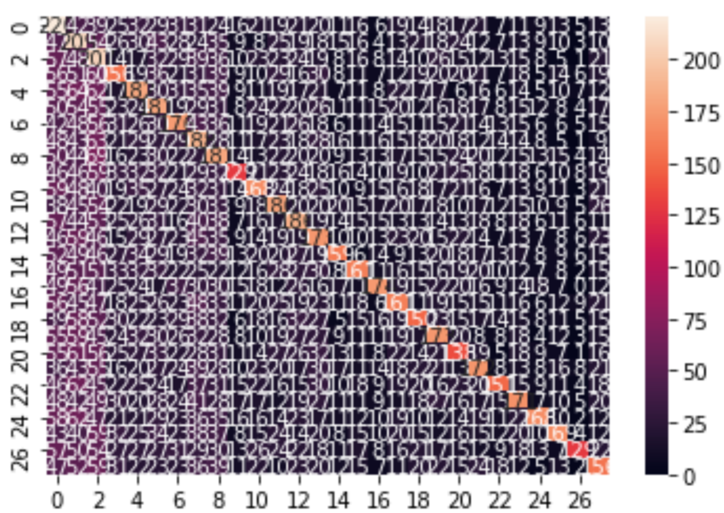

Figure 3: Confusion matrix for prediction

1.4 Gender Debiasing

We experimented with different threshold to achieve the highest model fairness without loosing f1 score. For threshold value of 4 standard deviation, best result was achieved. Another interesting finding from experimenting with different threshold is, as the threshold is pushed towards the boundary (at the point when f1-score starts to drop), there is a gain in f1-score until the boundary is violated. It turns out that there are some common features that negatively impacted during both gender and job classification and elimination of those common words keep happening until the boundary is violated. Trade-off between achieving the fairness and f1 score for different threshold is illustrated in Figure 4.

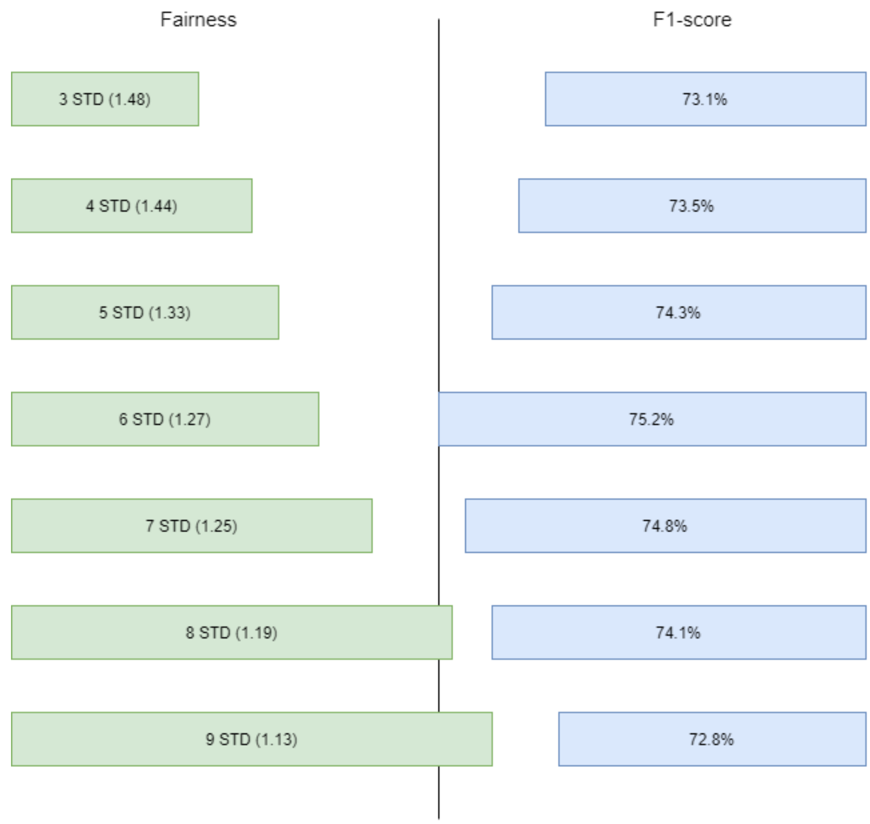

Figure 4: Fairness vs F1 score

2 Conclusion

Working over text data with probabilistic models is an interesting topic for us. Because of each document or observation consist of many words, it could come classify to any kind of class with respect to probabilities. We tried to overcome this different probabilities and find the best fitted class on each document with using Multinomial Naive Bayes, Latent Dirichlet Allocation and Gibbs Sampling.