Project Report in Text Mining

Attitude Analysis towards Immigrants among Swedish Unions

Jun Li

LiU ID: junli559
Course code: 732A92
Examiner: Marco Kuhlmann

Abstract

The report investigates the attitude towards immigrants in labor market among Swedish unions, through analyzing documents from 12 Swedish unions' press conference 2010-2020. A POS tagging model configured with pretrained word embeddings from FastText is employed to train on the dataset over different time span and unions. Relative norm distances between neutral and target word lists for Swedish and immigrants oriented groups are extracted from the retrained word embeddings. By analyzing the values of relative norm distances through years and unions, a relatively negative attitude towards immigrants among Swedish unions is detected in 2016, and LO Centralt shows a less positive attitude than other unions. Taking limited dataset and resources into account, improvements are possible for further research.

1 Introduction

Text mining (TM) has been one of widely applied technology in many scenarios for recent years, such as speech recognition and NPL etc. Due to rapid modern pace, people have not much time to digest a long article or newspaper as they did before. TM casts a light over this problem and offers possibility to process mass of text rapidly and effectively. At the same time, "*migrationsverket*" and immigrants have become one of the hot topics nowadays in Sweden, in particular after the refugee crisis in 2015 and even worse after COVID19 out-breaking. How the immigrants fit in Swedish labor market is one effective mirror to reflect this particular social field.

The report investigates the attitude towards immigrants in labor market among Swedish unions, through analyzing documents from 12 Swedish unions' press conference 2010-2020. In particular, a pretrained word embedding matrix by *FastText*[1] with common lemmatized words as in press documents after processing is inserted into a POS[2] classification model which is trained with sentences from press conferences. Together with neutral word list and target word groups, the retrained embeddings upon specific sentences is adopted to calculate the relative norm distances upon which the social science analysis is further developed.

[1] Word embeddings obtained at the link https://fasttext.cc/docs/en/crawl-vectors.html, which is trained on Wikipedia, having dimension 2,000,000*300 and powered by Facebook group

[2] Part-Of-Speech, is a tagging method for classifying words of a given corpus into corresponding part of context

The objective is to examine the trend of attitude climate through years and compare attitude states among unions. In order to evaluate this study in good manner, a pre-assessment over the situation is conducted: firstly, as known to all, there was the refugee crisis across whole Europe even in Sweden. It is complicated to conclude how the Swedish society responds to the impact, but generally there is a negative sentiment towards immigrants and refugees in large scope in those years. Secondly, it shows in 2002 that *Hotell- o Restaurangfacket* union has members with immigrant background[3] of around 36%, which is the largest proportion of all unions, followed by *Livsmedelsarbetarförbundet* with 29%, *Industrifacket* with 29% and *Metall*[4] with 26% [1]. Therefore it is assumed that these unions should bear relatively more positive attitude towards immigrants in Swedish labor market than other unions.

2 Theory

2.1 Relative norm distance

Nikhil Garg *et al.* has presented a systematic method of extracting information from trained embeddings by means of building words list for neutral and target groups, and calculating the relative norm distances RND (as in equation 1). The more positive/negative the RND is, the more associated the neutral words are toward group I/S [2], i.e. the Swedish work labor market or the specific union is more/less welcoming the immigrants.

$$rnd = \frac{1}{len(N)} \sum_{n \in N} \parallel n - S \parallel_2 - \parallel n - I \parallel_2 \qquad \text{(equation 1)}$$

where *rnd* is relative norm distance, *N* is the neutral word list, *S* and *I* are the average vector for Swedish respective immigrants word lists.

The project builds word lists from shared vocabulary by both pretrained embedding and dataset. Due to sparsity of *Swedish-oriented* words, most words in S-list are the traditionally unique Swedish names.

N-list=['arbetsmarknad','jobb','chans','tjänst','yrke',
'arbetsförmedling','arbetstagare','lön','intervju','anställning']
S-list=['svensk','medborgare','inrikes','Tomas','Björn','Lena','anna','Stefan',
'Karin','Sara','Malin','Andreas','Jan','Johansson','Maria','Erik','Sven',
'Peter','Carl','Mikael', 'Elisabeth','Eva','Kristina','Birgitta']

[3] Hereby both those born in and having at least one parent from foreign countries, i.e. the second generation immigrants

[4] *Industrifacket* and *Metall* merged into a new union *IF Metall* in 2005

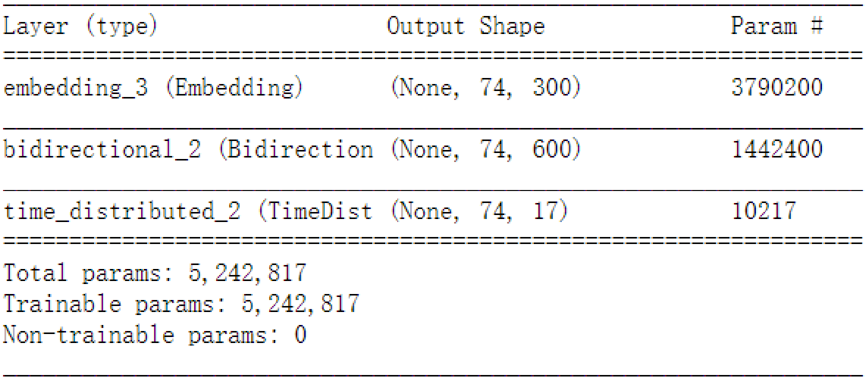
I-list=['invandrare','flykting','migration','visum','asyl','utomlands',
'arbetstillstånd','migrationsverket','asylsökande',
'arbetskraftsinvandring']

2.2 Transfer learning

Transfer learning is widely used in convolutional neural network training, which builds model and train parameters based on given trained model, in order to sort of transfer knowledge from other relevant model or public dataset to your own problem. This is usually suitable when there is small dataset in domain or to save training time [3]. Since the project has limited data size, a pre-trained word-embedding on Wikipedia is adopted.

2.3 Model

POS tagger is one popular method to build model and train word embeddings. In this project, the following model is employed, which consists of a embedding layer of size 74*300, a bidirectional LSTM layer of size 74*600 and a softmax layer of size 74*17 (74 is the max length of the sentences in dataset, and 300 is the length of word vectors). Besides, Adam optimizer is configured.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 74, 300)           3790200
_________________________________________________________________
bidirectional_2 (Bidirection (None, 74, 600)           1442400
_________________________________________________________________
time_distributed_2 (TimeDist (None, 74, 17)            10217
=================================================================
Total params: 5,242,817
Trainable params: 5,242,817
Non-trainable params: 0
_________________________________________________________________
```

Figure 1: Model summary

3 Data

3.1 Raw data

Raw data is kindly shared by Mr. Aliaksei Kazlou, the research fellow and his team at Department of Management and Engineering (IEI), Linköping University. There are 483 word documents of press conferences from 12 largest unions (as below) in Sweden in year 2010-2020. After processing the documents are converted and saved in a *.json* data-frame comprised by 15395 rows and 4 columns {*union, year, title, sentence*}.

* *A, LO centralt*
* *HR, Hotell- o resturangfacket*
* *H, Handelsanställdas förbund*
* *L, Livsmedelsarbetarförbundet*
* *K, Kommunalarbetarförbundet*
* *DA, IF Metall*
* *FF, Fastighetsansälldas förbund*
* *BA, Byggnadsarbetarförbundet*
* *E, Elektrikerförbundet*
* *M, Målarna*
* *S, Seko*
* *T, Transportarbetarförbundet*

3.2 Pretrained word embeddings

A word embedding pretrained on Common Crawl and Wikipedia is downloaded from FastText, which has 2 million unlemmatized word vectors of dimension 300.

4 Methods

The study is carried out in mainly 5 steps as follows[5]:

4.1 Prepare word embeddings

- tokenize 15395 sentences from dataset into lemmatized words and build a vocabulary of 12633 unique words, exclusive *'unk'*
- since the words in pretrained embedding are unlemmatized, loops are run to extract the lemmatized words shared by both embedding and dataset. There are 10939 common words, exclusive *'unk'*.
- only the vectors of common words are retained and will be inserted into the training model. For those words in pretrained embedding which share the same lemmatized word, such as "*blir*", "*blev*" having the same lemmatized word "*bli*", the average vector should be the case. Hence the initial embedding should be in shape (12634*300), inclusive '*unk*'

4.2 Model preparation

- divide processed dataset into training/validation/test datasets, including 0.7/0.15/0.15 sentences
- *Stanza*[6] is adopted to categorize words in each sentence into POS tags, which is considered as best substitutes for gold standard by

[5] Scripts for project are available at the GitHub link https://github.com/yewei369/TM.SU

[6] Stanza is created by Stanford NLP group, empowered by pretrained neural models supporting 66 human languages

author. And these categorized tags should be the targets in model training. Meanwhile, a list of unique tags for dataset is built

- convert each word and tag into unique integers, [0-12633] for words inclusive '*unk*' and [0-15] for tags
- tokenize and POS[7] 10776 sentences in training data, and pad each tokenized sentence with length less than 74[8] with special symbol from the beginning. Then encode the words and tags with unique integers. Hence, input with shape of (10776*74) and target with shape (10776*74*17)[9] has been built

4.3 Model selection

- configure hyperparameters *batch_size* as 32 and maximal *epochs* as 10, train and validate on 2309 sentences in validation set, select the best model with optimal epoch number which gives the highest accuracy or lowest loss value
- evaluate the selected model on test set with 2310 sentences and obtain the approximate model accuracy

4.4 Model training

- retrain on whole data set split on years, hereby for each year [2011-2019] the split dataset should contain the previous and coming years in order to reduce short-term data bias, for example, the trained embedding for year 2011 is actually trained over data for [2010-2012]
- retrain on whole data split on unions and obtain the corresponding embeddings for each union

4.5 RND analysis

- construct neutral word list and word lists oriented for Swedish and immigrants groups
- retrieve vectors for those words from the specified trained embeddings
- calculate RND for years and unions according to equation 1

5 Results

Restricted by the data size, the trained word embeddings vary for each training results, which should be explained by the random initialization for parameters in LSTM and *softmax* layers, which in return are incorporated

[7] Hereby POS means to categorize words into POS tags in each sentence
[8] 74 is the max length of 10776 sentences
[9] 17 is 16 tags plus one special tag for the padded symbols

when calculating the gradient descents of embeddings using backpropagation. Nevertheless, the training results show similar patterns for trend over years (figure 3), and differentiable characteristic among unions (figure 4).

5.1 Model validation

After training the model with 10 epochs, the training loss/accuracy is getting lower/higher all the way while validation loss/accuracy hits the min/max value after 3 epochs. Therefore epoch number is configured as 3. Then the selected model is evaluated on the test dataset and presents test loss of 0.0226 and Test accuracy of 0.9927.

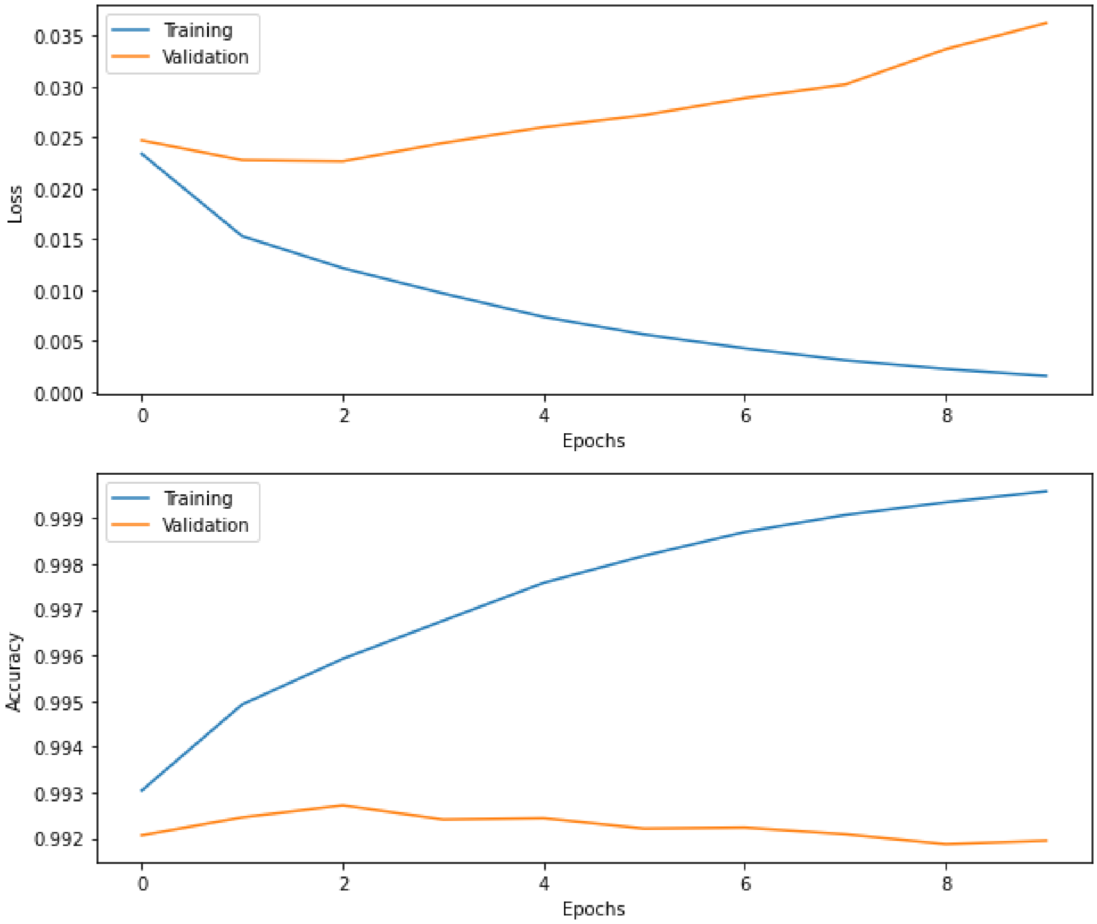


Figure 2: Model selection

5.2 RND over years

2 of the training results are presented as below, where table 1 summarizes the training performance and figure 3 for RND trend.

Table 1: Training result over years

Year	Sample size	Loss	Accuracy
2011	1745	0.2262/0.2257	0.9305/0.9315
2012	2374	0.1249/0.1069	0.9681/0.9731

2013	2703	0.0852/0.0892	0.9801/0.9791
2014	3748	0.0329/0.0299	0.9916/0.9921
2015	4569	0.0265/0.0253	0.9928/0.9930
2016	6243	0.0199/0.0206	0.9941/0.9939
2017	7344	0.0175/0.0174	0.9947/0.9947
2018	6925	0.0187/0.0186	0.9944/0.9945
2019	5179	0.0226/0.0225	0.9936/0.9936

* loss and accuracy are in format [first/second training]

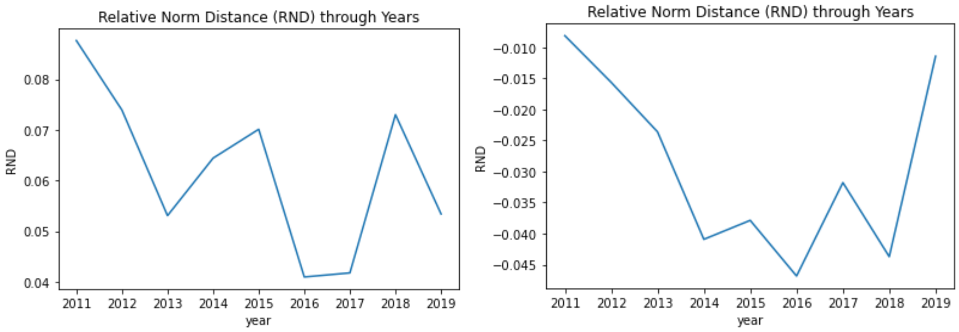


Figure 3: RND over years

5.3 RND over unions

2 of the training results are presented as below, where table 2 summarizes the training performance and figure 4 for RND characteristics.

Table 2: Training result among unions

Union	Sample size	Loss	Accuracy
HR	1370	0.2646/0.2666	0.9184/0.9147
M	362	0.4043/0.3983	0.8790/0.8790
L	1126	0.3098/0.3010	0.9008/0.9051
DA	735	0.3301/0.3303	0.8931/0.8931
FF	959	0.3428/0.3406	0.8849/0.8861
A	3834	0.0322/0.0322	0.9914/0.9916
E	654	0.3629/0.3598	0.8823/0.8832
H	772	0.3188/0.3190	0.8961/0.8956
BA	1380	0.2937/0.2774	0.9067/0.9137
T	1713	0.2249/0.2135	0.9296/0.9352
K	2097	0.1618/0.1602	0.9552/0.9569
S	393	0.3944/0.3993	0.8777/0.8777

* loss and accuracy are in format [first/second training]

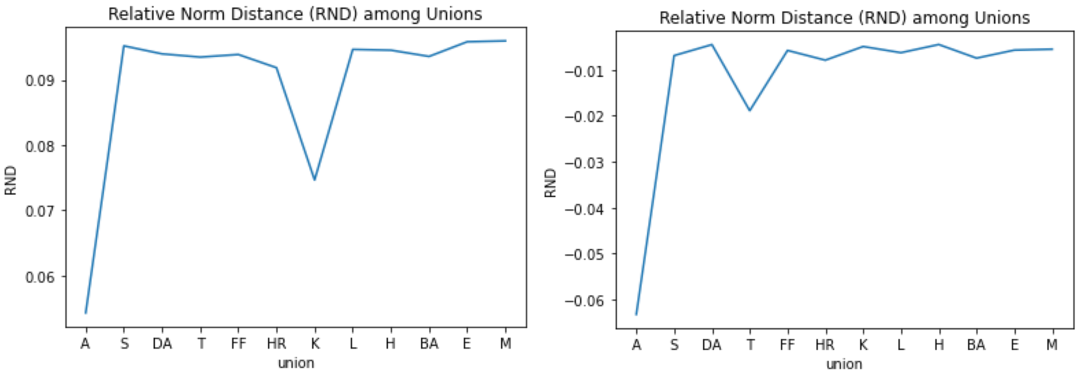


Figure 4: RND among unions

6 Discussion

6.1 Result analysis

The selected model seems to present high quality, since it reaches test accuracy of 0.9927. When the model get retrained on different year spans, it presents high accuracy almost 99% for most years. The same case does not repeat when training on unions, the accuracy are mostly around 87-90%, much lower than on years, which suggests that the quality of trained word embeddings on years are more stable than on unions.

Both RND trend over years shows the lowest value for year 2016, just after the refugee crisis in 2015, which reflects that in 2016 the Swedish labor market is more associated with native group. In other words, the immigrants were less welcomed in 2016, which is consistent with first hypothesis. That could be explained by the enormous impact of crisis on Swedish society in 2015 and a derived negative attitude towards immigrants got stimulated in 2016.

However, the RND values among most of the unions are at the same level except *LO Centralt* stands out with significantly low value in both training followed by *Kommunalarbetarförbundet* in the first training and *Transportarbetarförbundet* in the second, which indicates that *LO Centralt* tends to associate work market more with natives and less welcomes the immigrants. While *Hotell- o Restaurangfacket*, *Livsmedelsarbetarförbundet* and *IF Metall* do not show a significantly high value among all, which is not consistent with the second hypothesis. That could be explained by limited data size and training round, and latent climate change in labor market after 2002.

6.2 Contribution and limitations

Text analysis regarding immigrants in Sweden is absolutely new academic field, therefore it is difficult to find relevant literatures or gold-standard datasets. When it comes to attitude among Swedish labor market, the scope and resources become even limited. On the other hand, it makes this project innovative to some extent.

However, there are several inherent limitations in the project approach which could be improved in further study. Firstly, the data size is quite small comparing with the model complexity, even it is a simplified model adopted in this report; secondly, widely accepted ground dataset is absent; thirdly, POS model could be improved by skip-gram, Word2Vec, negative sampling or GloVe which should capture context information more efficiently; fourthly, as mentioned in results, the trained embeddings seem not stable for each training. Due to limited time span for this project, only 2 training are presented. But this problem should be mitigated through expanding the data size, or multiple rounds of training and get the average of word embeddings; last but not least, due to insufficient Swedish language knowledge, the neutral and target word lists could be rebuilt more professionally.

7 Conclusions

Classical NPL model employed in the report is verified to capture the context meaning, and generate decent semantic analysis to identify the latent negative attitude towards immigrants among unions especially in 2016, and referable climate map among Swedish unions. The project preliminarily explores the possibilities in this field, further research could be developed upon it.

8 References

[1] Kjellberg, Anders. "Ett nytt fackligt landskap - i Sverige och utomlands”, *Arkiv för studier i arbetarrörelsens historia*. 2002, nr 86-87, p.44-96

[2] Nikhil Garga, Londa Schiebingerb, Dan Jurafskyc, James Zoue. ”Word embeddings quantify 100 years of gender and ethnic stereotypes”, *PNAS*. 2018, Vol. 115, E3635-44

[3] Andreaw Ng. Deep Learning Specialization. Coursera.org