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 CBOW model configured with hierarchical softmax 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 trained 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 2015 and 2016, and *Fastighetsansälldas förbund* 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 NLP 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 CBOW[1] model configured with hierarchical softmax is trained by sentences from press conferences which consists of lemmatized words. Together with neutral word list and target word groups, the trained embeddings upon specific sentences is adopted to calculate the relative norm distances over which the social science analysis is further developed.

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,

[1] CBOW (continuous bag-of-words) is an architecture using single-hidden-layer neural network to predict target words with neighboring words, which is one typical approach of training word embeddings

as known to all, there was the refugee crisis across whole Europe even in Sweden around 2015. 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[2] of around 36%, which is the largest proportion of all unions, followed by *Livsmedelsarbetarförbundet* with 29%, *Industrifacket* with 29% and *Metall*[3] 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 built vocabulary by tokenized sentences of press conferences. Due to sparsity of *Swedish-oriented* words, most words in S-list are the traditionally unique Swedish names, which is inspired by Nikhil Garg *et al* 2018, where classical names of minority adopted to represent different ethnic groups.

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']

I-list=['invandrare','flykting','migration','visum','asyl','utomlands',

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

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

'arbetstillstånd','migrationsverket','asylsökande',

'arbetskraftsinvandring', 'nyanlända']

2.2 Model

Skip-gram and CBOW are the two dominant methods to train word embeddings, both using single-hidden-layer neural network to predict target by neighboring words. Since skip-gram predicts one target given one neighboring word and requires more data for a well-trained embedding, CBOW from *Word2Vec* in *Gensim* is employed in the report. In detail, model parameters are configured as:

window=5, min_count=5, size=100, sg=0, hs=1

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

4 Methods

The study is carried out in mainly 5 steps as follows[4]:

4.1 Prepare data

- extract and structure words from raw documents
- tokenize 15395 sentences from dataset into lemmatized words

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

4.2 Model training

- train 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]
- train on whole data split on unions and obtain the corresponding embeddings for each union, by which a particular vector space is draw from and meanwhile depicts the specified context of a union. Out of this space, the RND between Swedish and immigrants groups are supposed to capture the encoded distances, i.e. the attitude differences
- 100 training sessions are implemented for each variable including years and unions

4.3 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
- within each dimension (year and union), t-test is run between all variable pairs to check if their RNDs are significantly different

5 Results

5.1 RND over years

T-statistic values for RNDs in each pair of years are all approximately zeros, i.e. with 95% confidence level it is concluded that RNDs in all years are significantly different. Figure 1 shows the RND trend through years, solid line is the mean RND and shaded area the 95% quantile of 100 training results.

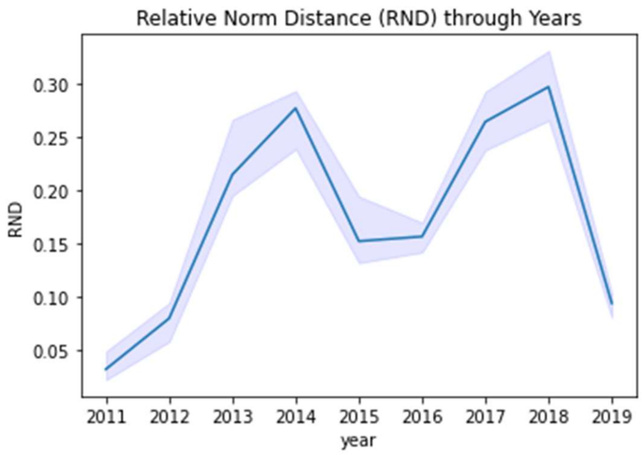

Figure 1: RND over years

5.3 RND over unions

T-statistic values for RNDs in each pair of unions are all approximately zeros, i.e. with 95% confidence level it is concluded that RNDs in all years are significantly different. Figure 2 shows the RND among unions, solid line is the mean RND and shaded area the 95% quantile of 100 training results.

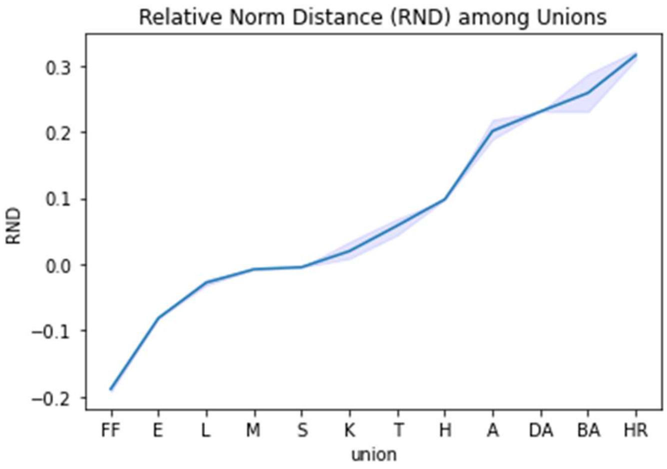

Figure 2: RND among unions

6 Discussion

6.1 Result analysis

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

However, the RND among unions show that *FF, E* and *L* possess lower values, which indicates that they tend to associate work market more with natives and less welcomes the immigrants. While *HR, BA* and *DA* top the list, which is 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 problem complexity; secondly, widely accepted ground dataset is absent for straight sentiment classification; thirdly, due to insufficient Swedish language knowledge, the neutral and target word lists could be rebuilt more precisely.

7 Conclusions

Classical NLP 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 2015 and 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 Garg, Londa Schiebinger, Dan Jurafsky, James Zou. ”Word embeddings quantify 100 years of gender and ethnic stereotypes”, *PNAS*. 2018, Vol. 115, E3635-44

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