IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Getting insight of employment vulnerability from online news: a case study in Indonesia[1]

Nursidik Heru Praptono and Alvin Andhika Zulen,
Bank Indonesia

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Getting Insight of Employment Vulnerability from Online News: A case study in Indonesia

Nursidik Heru Praptono[1], Alvin Andhika Zulen[2]

Abstract

As mass media nowadays moves into online platform, exploring textual information related to economic condition from online news is becoming computationally straightforward and is an interesting opportunity. We develop an online news-based indicator in order to help getting insight on the condition of employment vulnerability. A large number of online high frequency news captured from various news websites are then utilised to construct such indicator. Our finding is that some simple inference models for text classification combined with index calculation gives a promising information that strongly reflects the rate or the risk of being unemployed, given a certain time and a certain sector. The indicator built confirms other related indicators such as consumer confidence index from our consumer survey, indicator of job vacancy in Indonesia, and national statistics office data, represented by some number of Pearson correlation values. In addition to that, we also demonstrate that during the pandemic of COVID-19 in 2020, this indicator showed a sharp inclining curve especially in the 2nd quarter, reflecting that during this period so many labours are in a very high risk. We suggest that this indicator can be potential to support central bank policies, either as a leading indicator or as a quick alternative of survey based indicator.

Keywords: text mining, machine learning, employment vulnerability, unemployment rate, online news.

JEL classification: C38, C55, C11, E24, E27

[1] Department of Statistics, Bank Indonesia. email: nursidik_hp@bi.go.id (corresponding author)

[2] Department of Statistics, Bank Indonesia, email: alvin_az@bi.go.id

Contents

1. Background

The importance of employment vulnerability becomes obvious as the condition of massive unemployment can affect many aspects related to the macro-economy as a whole (IMF (2010)). The systemic financial stability for example, can be affected by the increase of unemployment (Hada et al. (2020)). The ones' ability of purchasing becomes relatively lower if the unemployment rate is high. The risk of increasing non-performing loan (NPL) would also become unavoidable. It is thus a serious concern for macroprudential policy makers to consider pay attention on employment vulnerability as it can yield any further serious systemic risks.

The unemployment and its rate is still a global issue at least within the last 4 years, with various emphasises. The International Labour Organisation (ILO) in 2018, stated that global unemployment remains elevated by more 190 million, while the vulnerable employment was recorded still on the rise (ILO (2018)). In 2019 ILO also reported that decent work deficits were widespread (ILO (2019)). The trend in 2020 stated that the total labour underutilisation was doubled as high as unemployment and about 8.8 percent of total working hour were lost -- of the COVID-19 pandemic effect -- despite about 30 million of new jobs were created (ILO (2020)). Finally, in 2021 ILO (2021) stated that the huge impact of COVID-19 proliferated so many aspects including long term unemployment issue due to economy recovery. Given those evidences, monitoring the condition of unemployment and its potential is important to conduct in order to see the country's economic health.

In the recent years, we have also witnessed that the access for information has obviously becomes easy as we are in the digitalisation era. This phenomenon enables us to access some insight from high resolution of data. Online news textual data is not an exception. Research have been conducted to investigate some information about e.g. economic related from textual information such as news or social media.

Through this paper, we propose a methodology on investigating the employment vulnerability index utilising online news data. The organisation of this paper is as follows: Section 2 describes the literature review related to the employment vulnerability/unemployment rate and some research related to text analytics utilisation to support economic/macroeconomic related-use cases. Section 3 describes our proposed methodology, while in Section 4 we discuss the results of our proposed methodology. Finally, we conclude our works and discuss some further directions in Section 5.

2. Literature Review

The employment vulnerability, in general, is a condition when an employee (or a group of employees) has a tendency that they are in the lack of decent working condition, lack of adequate social security, and shut 'voice' out through a representative board or similar related organisation (Johnson (2010)). This kind of employee typically either belongs to own-account worker and/or contributing family worker. The employment vulnerability however has a close relation to the unemployment rate, although the latter is rather reflecting a condition/degree of unemployment. Identifying employment vulnerability could help anticipating massive unemployment rate.

Indonesia, as one of the most populous countries in the world is not without exception in having unemployment issue. The level of unemployment rate in Indonesia is still relatively higher than most ASEAN countries (OECD, (2020)). Our national statistics office (BPS) recorded that by February 2021, the open unemployment rate in Indonesia reached about 6,26%, lower than the previous 6 month (7,07% in Augsut 2020). The method conducted by BPS in order to produce such information is based on survey called Survei Angkatan Kerja Nasional (Sakernas) -- national labour force survey, whereas the respondents are in the productive age (BPS (2021)). However, as we may see, the obvious limitation on measuring the unemployment rate is that it is based on the 6-monthly basis. For e.g. macroprudential policy makers that need quicker analysis, an alternative way to gather the likely related information of unemployment in high frequency is thus required. One of the alternative data source to explore is online news, that is textual data that can be produced in near real time.

The research related to the utilisation of textual analytics on economic news have recently been conducted. Baker et al. (2016) in their work utilised news data from 10 leading US newspapers in order to construct the economic policy uncertainty index. The method the implemented is by providing a set of terms that reflect the economic uncertainty. The index is then constructed based on the classified news data compared to the available article data. Generally, the text analysis conducted in this case is rather of deterministic approach.

Another work on economic-related text analytics that utilising machine learning is the work of Tobback et al. (2017). Using textual data from media, their experiment result showed that a hawkish-dovish degree related to the central bank's communication measurement is better constructed by leveraging support vector machine. Moreover, in their report, Latent Dirichlet Allocation (LDA) is then utilised to identify the topic, on to which certain degree of communication measured.

The use of social media as the source of economic textual data has also been conducted, for example by Bollen et al. (2011). The stock markets are predicted by analysing the mood information inferred from twitter data. It is found that the prediction accuracy reached about 87,6% on the prediction given their prediction model setup.

Another experiments on social media data that is related to the issue is the work by Antenucci et al. (2014). The twitter data is used to see the labour market flows by creating jobs related indexes, ranging from "job loss", "job search", and "job posting". The result demonstrates that such constructed index, given the setting, can be used as the consideration related to insurance policies/support.

Recently, research work on more related to the employment vulnerability from the textual data has been conducted by Bailliu et al. (2018). They developed an index called Chinese Labour Market Conditions Index (LMCI) in order to measure the condition of labour market in China. This work utilises Support Vector Machine (SVM) to classify the newspaper article. Furthermore, their result also suggested that having setup as defined, the LMCI can be used in forecasting the labour market condition.

Having the discussion as above, we then generally summarise two main research questions to answer through our experiment:

1. How to classify the text with the limitation of training data while having prior knowledge?

Of the textual mentioned above, we however still have a big challenge when utilising textual data using either deterministic (rule based with some predefined keywords) or machine learning models. When using rule based method, the model's generalisation ability is solely based on human prior knowledge. On the other hand, machine learning model relies heavily on the data and thus can suffer from low performance when the quantity and the quality of the training data is insufficient.

2. Is there any better suggested method to construct employment vulnerability index with fine grained time basis (near realtime)?

To the best of our knowledge, there is currently no official survey based data that we can refer related to the employment vulnerability index in Indonesia. Even if it does exist -- say informally--, the cost could be very expensive. Moreover, it may suffer from non-trivial limitation such as subjective bias and less seamless information captured.

3. Methodology

3.1. Data

3.1.1. News Article

We utilise news articles for our main source of data, from more than 30 various domestic online news portals from January 1998 to August 2021. The average of total article is about 850 per day and about 27.000 per month. In order to demonstrate our methodology, we define some setups as the following:

1. **News filtering**: Of those the whole news, we filter the articles for where there is at least one sentence that contains any keywords related to unemployment. Thus we first perform sentence tokenisation onto each article. If an article does not contain any such keywords, then we simply flag this article as not indicating any employment vulnerability.

2. **Data Annotation**: Of the filtered articles as described in point (1) above, we subsample the data per month such that we obtain a 10% on each month. The time interval of this data ranges from January 2020 up to December 2020. This filtered data are then pooled out for annotation process by human, i.e. to flag whether an article is 0 or 1, based on the filtered sentence(s). Here label 1 means "the article indicates any information related to employment" where 0 means "the article does NOT indicate any information related to employment". The overall annotated data consists of 2979 of class 1 and 3167 of class 0. Of the number of records in class 1, we identify that there are 196 articles belong to manufacturing sector and 55 belong to service sector. We utilise this annotated data to construct and/or evaluate the text-classifier model.

3. **Full Data**: the full data for demonstrating the model on classifying the texts and constructing the index are of those articles from August 2018 to August 2021 (monthly).

3.1.2. Supporting Data for Index Evaluation

For the evaluation of our constructed index, we elaborate the job vacancy index data obtained by various online job bursaries. This data represents the number of job

vacancy within a certain period. The higher job vacancy index, the higher job offered on the market. The index is available up to monthly. Another data that we have is consumer confidence index (CCI) provided monthly from our Consumer Survey. We also utilised the GDP data by specific sector, provided by BPS and available quarterly.

3.2. Inference Models

In this part we will discuss some alternatives of the inference models and their properties. Given an x (article), the model is to predict its corresponding category (class) $\hat{y} \in \{0, 1\}$. First we describe the deterministic model, then model-from-data and finally suggest the inference model that incorporating prior knowledge into model-from-data. The idea behind our experiment is that we would like to utilise our prior knowledge, expecting that it will help the inference process. This is as an alternative when we have any limitation to access the training data, while still be able to make any room for prior knowledge for inference process.

3.2.1. Deterministic Model

The deterministic model we apply here leverages the rule based model and some selected keywords. The objective of the rule is basically to find any pattern indicating the evidence of unemployment within the selected article.

$$r(x) = \exists(\mathrm{kw}_{\mathrm{unemployment}}) \text{ in } x \bigwedge \nexists (\mathrm{kw}_{\mathrm{neg1}}, \mathrm{kw}_{\mathrm{unemployment}}) \text{ in } x \bigwedge \nexists(\mathrm{kw}_{\mathrm{unemployment}}, \mathrm{kw}_{\mathrm{neg2}}) \text{ in } x \quad (1)$$

Some keywords example used by Eq. 1 can be seen in Table 1 as the following.

Some keywords example and their English meanings — Table 1

Keywords Category	List of Keywords Example	Representative Meaning (English)	Notes
Keywords unemployment (kw_unemployment)	pemutusan hubungan kerja, phk, pemulangan, pengangguran, layoff, pemecatan, memecat, dipecat, pemulangan,...	≈ Fired-out, layoff	-
Keywords negation 1 (kw_neg1)	tidak	≈ no/there is/are/were no	Co-occurance: before kw_unemployment
Keywords negation 2 (kw_neg2)	menurun, turun, berkurang, melandai, menyusut, rendah,...	≈ decreasing, becomes lower.	Co-occurance: after kw_unemployment

For further utilisation in our experiment when elaborating with model-from-data (machine learning), we simply refer this deterministic model $r(x)$ as our "prior knowledge". This is because this model is simply as the representation of human knowledge.

This model in one hand is rather simple, rigid, and straightforward to apply. However, it is very sensitive to the change of characteristics of the data. The assumptions provided to construct the rule must capture all the possibilities of the

whole patterns in order to obtain best performance, and so do the keywords. Otherwise, if the model is too much simple, it may result so many unexpected noises.

3.2.2. Inference Model from Data

Another approach to classify the filtered text with more flexible way is to build a model-from-data (or in general term we may say "machine learning"). In our case we use Logistic Regression to demonstrate the inference model. Thus we have:

$$p(y = 1|x) = \sigma\big(f(x)\big) = \frac{1}{1 + e^{-f(x)}} \tag{2}$$

where $f(x) = w\phi(x)$ represents our linear function, with w is the models parameter and ϕ as basis function. In our case, we simply use simple polynomial basis function with degree $m = 1$, thus $\phi(x)$: $x \to x$.

Given the annotated dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ for $x \in \mathbb{R}^d$ and $y \in \{0,1\}$ the objective function is to minimise its negative log likelihood, that is

$$J = \sum_i \ln(1 + e^{(-(2y_i-1)f(x_i))}) \tag{3}$$

Having the formulation as defined above, the goal of our optimisation problem is thus to find w that minimise J on such setting. In our experiment, we demonstrate the estimation of the parameter by quasi-newton approximation with Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Note that before the article is used by the model, we perform feature extraction first by leveraging TFIDF[3]-bag of words feature extraction on the related sentences so that each article can be represented by a feature vector.

As it is of any general machine learning methods, this approach is more flexible in term of constructing knowledge as long as the samples obtained are representative enough to the whole population. However, when the number of the training data is rather limited or less representative to the population, the model learns insufficiently and perform badly on the prediction process.

3.2.3. Incorporating the Prior Knowledge

We finally tried to incorporate our prior knowledge into the model-from-data for inference process. While straightforward, incorporating prior knowledge needs concise modelling so that the it is properly blended into the model-from-data. We adopt the work of Schapire et al. (2002) for where they incorporate prior knowledge into AdaBoost. However, in our case we use the simpler model Logistic Regression to demonstrate the idea while applying into our text classification problem.

First, we introduce a probability distribution that enables to represent our prior knowledge. We refer this prior probability as π_+, that is the probability that any x would belong to class 1. In general, the form of the prior distribution indeed may vary depending on ones' choice. It may be justified either by human belief or by another mechanism that might rely on the data first before learning. We define our prior knowledge distribution in this case as in Eq. 4 follows:

[3] TF = Term Frequency; IDF = Inverse Document Frequency

$$\pi_+ = p(y=1|x) = \begin{cases} 0.9, & \text{if } r(x) = 1 \text{ (True)} \\ 0.1, & \text{otherwise} \end{cases} \quad (4)$$

Here $r(x)$ is obtained by our deterministic model as described in 3.2.1 previously. Simply speaking, if any article follows the rule, then we inject the probability value as 0.9, while 0.1 otherwise.

This prior probability is then taken into the objective function when estimating the model's parameter. The objective function on Eq. 3 thus becomes:

$$J = \sum_i \ln\left(1 + e^{(-(2y_i-1)f(x_i))}\right) + \underbrace{\eta D_{KL}(\pi_+(x_i)||\sigma(f(x_i)))}_{\text{control on prior information}} \quad (5)$$

The second part of the equation represents the control on our prior knowledge. It is quantified by Kullback-Leibler Divergence between π_+ (our prior knowledge representation) and "information from training data", $f(x)$. The value of η controls the importance of our prior information. Here we set its value into 1 and we leave as it is since in this experiment, it is not our main focus. In estimating the parameters, we apply the quasi-newton approximation with BFGS algorithm to demonstrate the learning process. As in the previous section on model-from-data, we first convert our article into feature vector by leveraging TFIDF bag-of-words feature extraction.

Once the parameter is estimated, we then introduce the prior term h_0 in order to enable the blending process into the model-from-data's term. This h_0 is obtained by the inverse of our prior probability π_+, so that:

$$h_0(x) = \sigma^{-1}(\pi_+(x)) = \ln\left(\frac{\pi_+(x)}{1-\pi_+(x)}\right) \quad (6)$$

The final prediction model is thus of the form of logistic function of the final blending $p(y=1|x) = \sigma(f^*(x))$ where $f^*(x) = f(x) + h_0(x)$ represents our final blending term and $\sigma(.)$ is the logistic function.

3.3. Index Construction

Once the full data have been categorised whether belong to 1 (article indicates any contains related to employment vulnerability) or 0 (article does not indicate), the index is then constructed. The equation below formulates how we construct the employment vulnerability index.

$$\text{idx}_t = \frac{\sum_i 1_{f(x_{i,t})=1}}{\text{N}_t} x \log \text{N}_t \quad (7)$$

Generally, the index is constructed based on the proportion of related articles among the overall articles within a time. Here $\sum_i \text{I}_{f(x_{i,t})==1}$ indicating the total number of related articles, while N_t represents the total of articles within a time t.

The log term in the Eq. 7 represents the importance, or the magnitude of the information. We realise that there may be a difference of the total articles between time. Thus, the more article within a particular time, the more important ratio should be considered. We understand that this approach might not be a perfect formulation but we fundamentally argue that it is of the most reasonable approach that we can utilise.

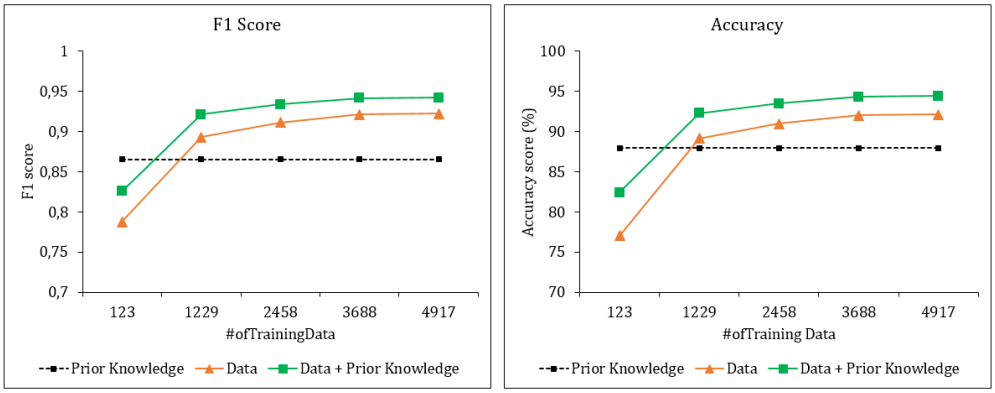
4. Result and Discussion

4.1. Evaluation on Text Classification Models

On experiment for text classification, we test each model into a fixed test data. As mentioned on the subsection 3.1.1 above, we have 6.146 annotated data points (2.979 of class 1 and 3.167 of class 0). Of those overall annotated data, we spare 20% as fixed testing data. The remaining portion is used for training experiments.

We compare three different approaches of inference models, which are deterministic/rule based (we note as "Prior Knowledge"), model-from-data/machine learning (we note as "data") and the incorporation of prior knowledge into model-from-data/machine learning (we note as "Data + Prior Knowledge"). We evaluate the model's performance with F1-score and accuracy score.

Experimental Results on Text Classification Model — Figure 1



a. F1 score of each model

b. Accuracy score of each model

The experiment for text classification can be seen as in Figure 1 above. Initially, when the size of training data is relatively small, the "data" model performed worst. We can see on Figure 1.a that our "prior knowledge"-- that is rule based -- is even still better than of "data" model on limited training data. As we increase gradually the size of training data, the F1-score is getting better. The "prior knowledge" only is constant, as it is a deterministic function.

It is can also clearly be seen that incorporating "prior knowledge" helps the performance of "data" models at any condition. On a small dataset, although below the rule based model, the performance when we incorporate prior knowledge help increase the model-from-data only. However, when there is sufficient amount of training data inference model that incorporating prior knowledge performs very well compared to model-from-data only or prior knowledge only.

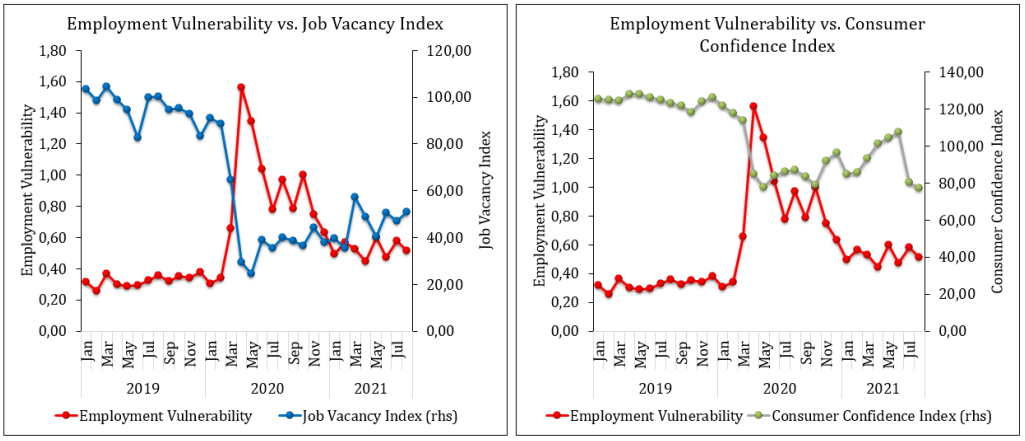
The accuracy of the model, on the other hand also show the similar trend, as can be seen in Figure 1.b. Initially, the model-from-data poses lowest accuracy on limited training data and while prior knowledge model is still the highest. Incorporating the prior knowledge, increase the accuracy although it is still lower than prior knowledge only. Once we have sufficient training data, the performance getting better when blending those both information for inference process.

4.2. Employment Vulnerability Index and its Evaluations

Once the articles have been classified, the employment vulnerability index is then straightforwardly constructed. Utilising Eq. 7, we plot our index based on timely basis. Figure 2 below shows the general monthly employment vulnerability index and its comparison with Job Vacancy Index and with consumer confidence index.

Employment vulnerability index vs. job vacancy index and consumer confidence index (CCI)

Figure 2



a. Employment vulnerability index vs. job vacancy index

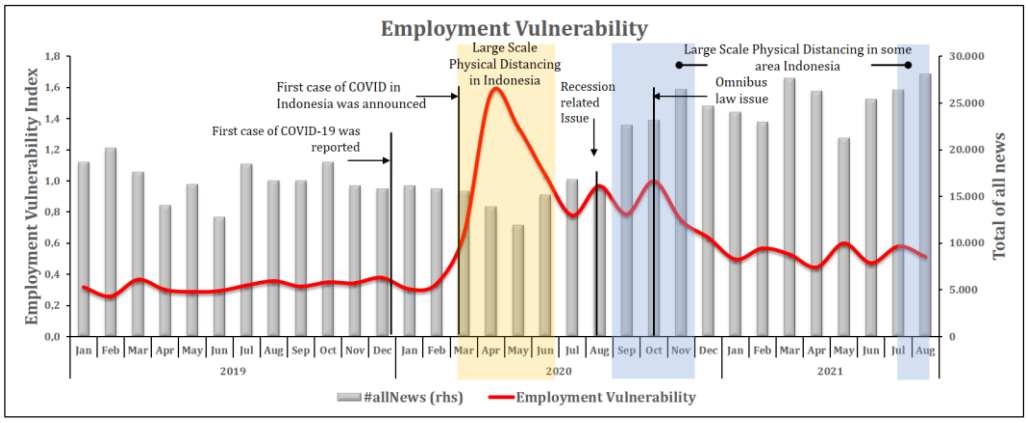
b. Employment vulnerability index vs. CCI

We found that generally the employment vulnerability index has the opposite direction to job vacancy index as we can see on Figure 2.a. above, having Pearson correlation of $\rho = -0.76$. In addition to that, its comparison to consumer confidence index also shows similar wise (as shown on Figure 2.b.), with the Pearson correlation of $\rho = -0.9$. The employment vulnerability index revealed the sharp inclining curve on the occurrence of the pandemic of COVID-19, especially since March 2020. During this period, so many labours are in a very high risk. This is due to the policies carried out by the government to anticipate the spread of COVID-19 by ruling strictly physical distancing related regulations on the whole country. The number of job vacancies drops dramatically in this period and so does the consumer's confidence.

We also can see that employment vulnerability index however is still relatively high at least until the end of 2020 compared to before. It is also found that there are some small fluctuating curves afterwards. We then performed the event analysis on the inferred articles as shown in Figure 3, mainly based on the content of the articles and related policies. We found that these small peaky curves indicate some events that can cause or have association with the employment vulnerability, such as the occurrence of recession issue, some demonstrations related to labour regularisation, and physical distancing on some particular area.

Event analysis on monthly employment vulnerability index

Figure 3

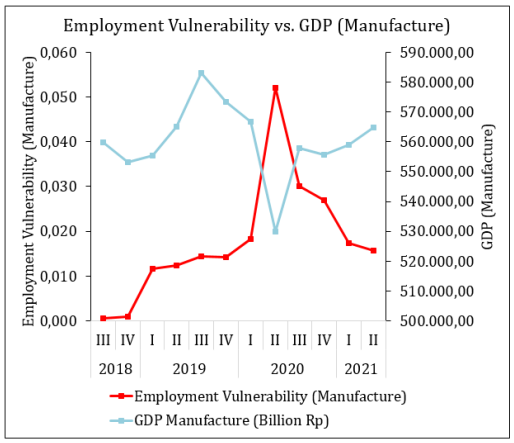


We then investigate the employment vulnerability on manufacturing sector and service sector, as they pose relatively significant sector on labour employment vulnerability. Most of the labours or employees are work on manufacture companies as well as service sector. We then compare such index with its GDP respectively.

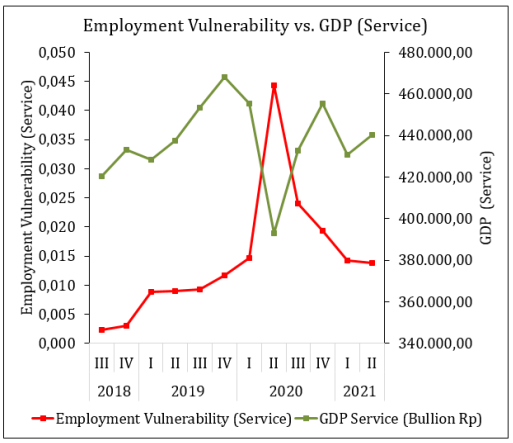
The GDP of manufacturing sector consist of those GDP from manufacturing industries. The service sector consists of those GDP from some subsectors such as transportation and warehousing, finance and administration, company service, educational-related service, social/society-related service, and other services.

Employment vulnerability index on manufacturing sector and service sector

Figure 4



a. Employment vulnerability index on manufacturing sector and its GDP respectively



b. employment vulnerability index on service sector and its GDP respectively

We found that the employment vulnerability index on manufacturing sector has strong Pearson correlation with its GDP in the opposite direction ($\rho = -0.8$). The GDP drops dramatically on 2nd quarter of 2020 as well as the occurrence of pandemic COVID-19 (see Figure 4.a). It is then increasing gradually as the employment vulnerability index decreasing due to some regulation adjustments and economic policies for facing the new normal.

In the service sector, such trend applies similarly (see Figure 4.b.). The employment vulnerability index have strong opposite direction with its GDP, with the Pearson correlation $\rho = -0.72$. The GDP curve also dropped dramatically on the 2nd quarter of 2020, as the so many employees are in very high risk during such condition. It then increased gradually as the employment vulnerability index decreased also together with some regulation adjustment and economic policies on the new normal era.

We have shown that some macroeconomic-related indicators confirmed the employment vulnerability index. Having text classification based index construction above helps policy maker get insight on what would be the condition of employment's vulnerability. Although we demonstrated and discussed monthly and quarterly time basis, the employment vulnerability index can however straightforwardly be presented in the daily basis timeframe. Such finer grained representation is useful as it can provide the information in near real time.

5. Conclusion & Future Directions

5.1. Conclusion

Towards this research we conducted a methodology for assessing the employment vulnerability from online news. We demonstrated on how we enable our prior knowledge to be incorporated into the model-from-data (machine learning model). We found that given limitation of training data, the incorporating prior knowledge helps the inference models when classifying the texts. On the other hand, the use of model-from-data can also help the rule-based model as it enables the construction of knowledge representation from data. Thus, we suggest that incorporating prior knowledge for such inference task is important to consider.

The index constructed generally confirmed by related indicators, including e.g. job vacancy, CCI, and GDP-per sector. This has been shown by relatively strong Pearson correlation on the demonstrated time range, either on general unemployment vulnerability index, or per sector unemployment vulnerability index. The index constructed from online news is a form of very high frequency data in that it can be obtained by daily. Thus, we also suggest that the proposed methodology can be used as a leading indicator or as a quick alternative to survey based index of employment vulnerability.

5.2. Future Directions

We notice that there still some improvements to do for our works. We highlight some future directions:

- Stratify the level of employment vulnerability index

Instead of binary classification, we suggest that the classification task is possible to expand to multi label classification representing the level of vulnerability. This is also pointed out by survey based EVI (Baum, S., & Mithcell, W., (2020)) in that they stratify the level into high, medium, low risk. The higher level should obviously be more concern for policy makers to do some essential and important decisions to prevent any worse condition.

- Enhancement on Spatial Information

The proposed methodology is solely based on the overall area on a state. However, it also may be important to localise the employment into some specific area. This is to help seeing the distribution of vulnerability by spatial information, either on binary categorisation, or multi-level/stratified categorisation.

- Enhancement on more sectors.

The two sectors discussed in this works are based on the majority of workers. However, it is also important to expand the sectors horizontally, or vertically (detailing into subsectors). However, this also depends on the purpose on how detailed information one should investigate, by sectors/subsectors.

References

Antenucci, D., Cafarella, M., Levenstein, C., M., Re., C., & Shapiro, M., D., (2014). Using Social Media to Measure Labor Market Flows. *NBER Working Paper No. 20010*.

Bailliu, J., Han, X. &, Kruger, M., (2018). Can media and text analytics provide insights into labour market conditions in China?. *Bank of International Settlements working paper*. URL: https://www.bis.org/ifc/publ/ifcb49_44.pdf

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics, 131*(4), 1593-1636.

Baum, S., & Mithcell, W., (2020). Employment Vulnerability Index (EVI) 3.0. Retrived from EVI's official website. URL: http://www.fullemployment.net/publications/reports/2020/EVI_3.0_Final_Report.pdf.

Bollen, J., Mao, H., & Xiao-Jun, Z. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science, 2*(1), 1-8.

Hada, T., Barbuta-Misu, T., Iuga, I. C., & Wainberg, D., (2020). Macroeconomic Determinants of Nonperforming Loans of Romanian Banks. *Sustainability 12* (7533), 1-19.

International Labour Organisation (2018). *World Employment Social Outlook*. Retrieved from official ILO's website. URL: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_615594.pdf

International Labour Organisation (2019). *World Employment Social Outlook*. Retrieved from official ILO's website. URL: https://www.ilo.org/global/research/global-reports/weso/2019/lang--en/index.htm

International Labour Organisation (2020). *World Employment Social Outlook*. Retrieved from official ILO's website. URL: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_734455.pdf

International Labour Organisation (2021). *World Employment Social Outlook*. Retrieved from official ILO's website . URL: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_795453.pdf

International Monetary Fund (2010). Unemployment dynamics during recessions and recoveries: Okun's law and beyond in World Economic Outlook (Chapter 3). pp. 1-39. URL: https://www.imf.org/~/media/Websites/IMF/imported-flagship-issues/external/pubs/ft/weo/2010/01/pdf/_c3pdf.ashx

Johnson, J. L., (2010), ILO Online: How do you define 'vulnerable employment?'. Retrieved from official interview with ILO chief of employment trends unit. URL: https://www.ilo.org/global/about-the-ilo/mission-and-objectives/features/WCMS_120470/lang--en/index.htm

OECD, (2020). Promoting Stronger Local Employment in Indonesia. Retrieved from OECD's official website. URL: https://www.oecd-ilibrary.org/sites/1f8c39b2-en/index.html?itemId=/content/component/1f8c39b2-en#chapter-d1e9121

Schapire, R., E., Rochery, M., Rahim, M., & Gupta, N., (2002). Incorporating Prior Knowledge into Boosting. Proceeding of the Nineteenth International Conference on Machine Learning, (pp. 538-545).

Statistics Indonesia (BPS), (2021). Tingkat Pengangguran Terbuka. Retrieved from BPS' official website. URL: https://www.bps.go.id/indicator/6/543/1/unemployment-rate-by-province.html.

Tobback, E., Nardelli, S., & Martens, D. (2017). Between Hawks and Doves. *NBER Working Paper No. 2085*.

BANK INDONESIA

Getting Insight of Employment Vulnerability from Online News: a Case Study in Indonesia

[1]Nursidik Heru Praptono, [2]Alvin Andhika Zulen

[1]nursidik_hp@bi.go.id
[2]alvin_az@bi.go.id

October 2021

BANK INDONESIA

Outline

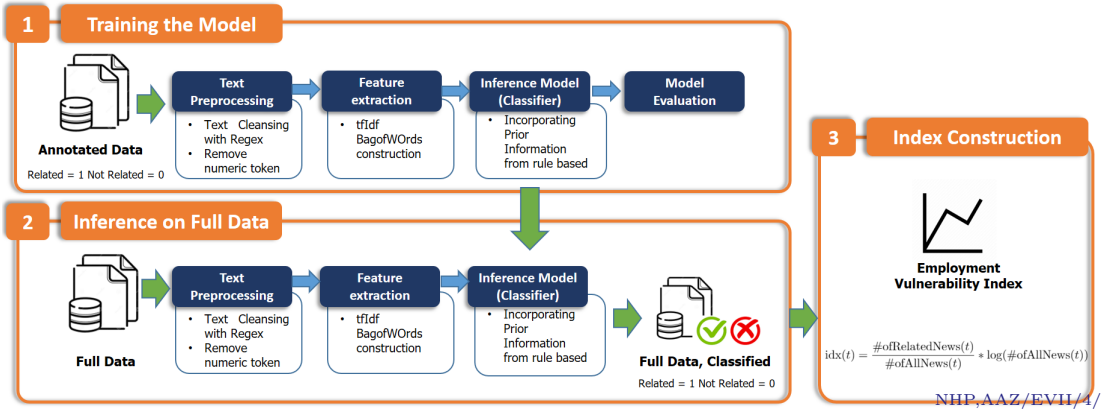
BANK INDONESIA

Background

- The unemployment rate is obviously an important factor that can reflects the health of the economy, e.g. systemic financial stability.
- Employment Vulnerability Index in Indonesia
  - Can have insight for macroprudential policy makers in order to make any further decision.
  - Currently there is no survey to construct the employment vulnerability index. Even if it exists, it may be expensive, subjective/biased and less seamless.
- Extraction economic-related information from news, e.g.:
  - Economic Policy Uncertainty (Baker et al, 2016)
  - Labour Market Condition Index, LMCI (Bailliu et al, 2018)
- Thus we introduce a methodology to enable employment vulnerability index, leveraging online news.
  - Intuitively, the more number of published articles related to e.g. unemployment, then there may potentially be more information that reflect the condition of employment vulnerability.

Overall Methodology

To construct the index, we first classify the text that possibly belongs to 1 (related) or 0 (not related). In general, our text classification methodology is rather straightforward as shown in part 1 and 2 of the diagram below. However we consider incorporating prior knowledge into our classifier in order to anticipate limited number of training data available. Once data is inferred, then the index is constructed (part 3).



Data & Prior Knowledge

Data

Overall data: Online domestic news data from more than 30 various news portal, from Jan 1998 to Aug 2021. Total average 850 news per day, and about 27.000 news per month.

- Annotated Data: 6146 (2979 of class 1, 3167 class 0), filtered by keywords related to unemployment. It is of sampled data from 10% randomly per month from year 2000 to 2020. Of the 2979 records of class 1, we identify 196 articles belong to manufacture sector and 55 articles belong to service sector.
- Full Data: All articles from 2018 to Aug 2021.

Prior knowledge setup: Rule and Keywords $r(x)$

We introduce rule and predefined keywords as a representation of our prior knowledge to classify the text. This model will then be used as our basic deterministic model.

$r(x) = \exists(\text{kw\_unemployment}) \text{ in } x \wedge \nexists(\text{kw\_neg1,kw\_unemployment}) \text{ in } x \wedge \nexists(\text{kw\_unemployment, kw\_neg2}) \text{ in } x$

list of keywords (their meaning in English):
kw_unemployment = {unemployment, layoff, fireout ....}, kw_neg1 = {not, avoid, reduce,...}, kw_neg2 = {decrease, step down,...}

Inference Model for Text Classification

Applying machine learning is challenging when there is limitation on accessing the training data. On the other hand, quite often we have intuition about something. Why not we elaborate our prior knowledge for the better inference?

- We adopt the methodology incorporating prior knowledge described by Schapire et al, 2002, but in our case we demonstrate the simpler model, Logistic Regression.
- **Objective Function**: Given the training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, and $y \in \{0, 1\}$. The objective function is to minimise negative log likelihood, controlled by the prior information.

$$J = \sum_i [\ln(1 + e^{(-(2y_i-1)f(x_i))}) + \underbrace{\eta D_{\mathrm{KL}}(\pi_+(x_i)||\sigma(f(x_i)))}_{\text{control on prior information}}]$$

where f is linear function, σ is the logistic function. Here π_+ is our "prior information" defined by:

$$\pi_+(x) = p(y = 1|x) = \begin{cases} 0.9; \text{ if } r(x) = 1 \text{ (related article, by rule-based model } r) \\ 0.1; \text{ otherwise} \end{cases}$$

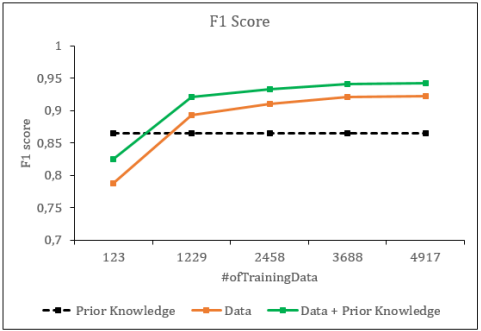
- **Inference Function**:

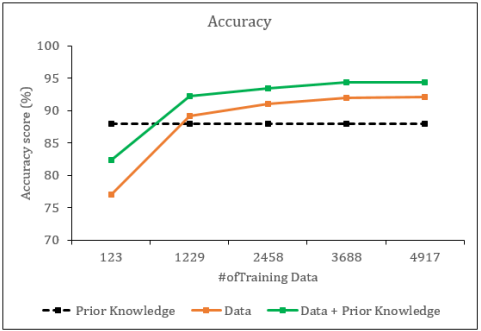
$$p(y = 1|x) = \sigma(f^*(x)); f^* = f + h_0$$

In this case, h_0 is the "prior term" defined as the inverse of logistic function of $\pi_+(x)$, that is $h_0(x) = \sigma^{-1}(\pi_+(x)) = \ln\left(\frac{\pi_+(x)}{1-\pi_+(x)}\right)$

Experimental Result

On a **fixed test data**, we compare three different approaches to classify the text. The model from data performs worst on small number of training data. As we increase the size of training data gradually the performance gets better. The model with prior knowledge only is constant, as it is a deterministic function leveraging rule based and predefined keywords. Incorporating prior knowledge into model-from-data's in this case helps in improving the performance, even when the number of training data is relatively small.

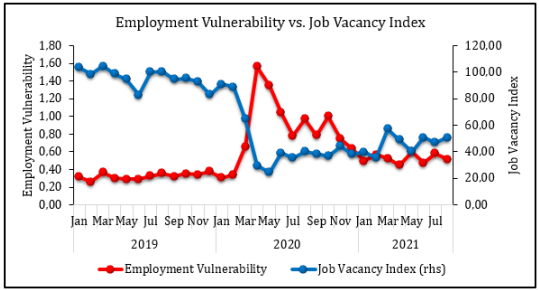


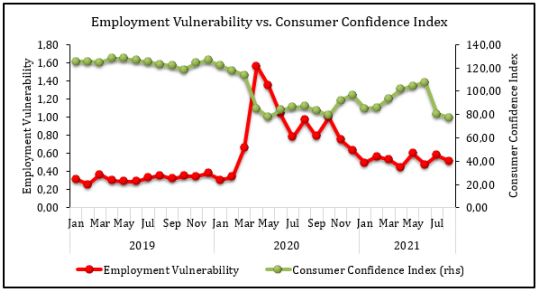


BANK INDONESIA

Employment Vulnerability Index

Once the articles are classified, we then construct the index. We found that generally the employment vulnerability index reveals the opposite direction with both the Job Vacancy Index ($\rho = -0.76$) and the Consumer Confidence Index ($\rho = -0.9$). The job vacancy index, describes the index of available job at a certain period. The consumer confidence index describes the consumer's confidence, obtained by survey. The curve shows an increase starting on March 2020 as in such period there were many actions to anticipate the COVID-19 pandemic.

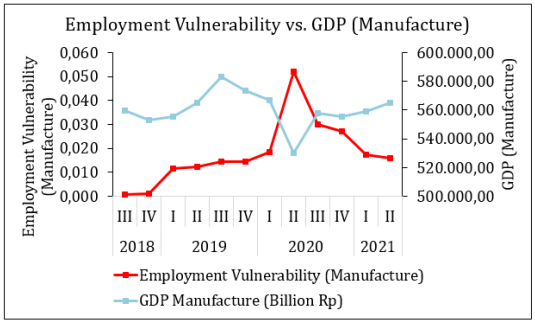


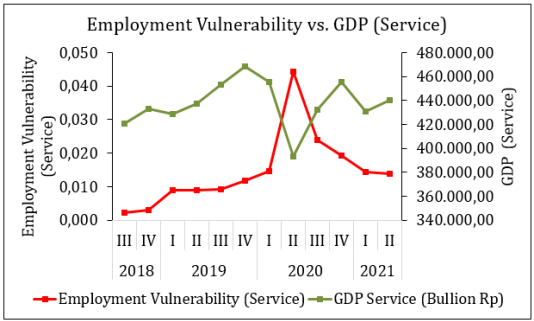


BANK INDONESIA

Employment Vulnerability Index - On Manufacture and Service Sector

Of the related news inferred by proposed approach, we also calculate the index per sector: manufacture and service. We found that there is relatively strong correlation (in opposite direction), on each sector with its GDP, $\rho = -0.8$ for Manufacture sector and $\rho = -0.72$ for Service sector respectively.





BANK INDONESIA

Conclusion and Future Directions

Conclusion

1. Incorporating prior knowledge into model-from-data can be helpful for text classification especially when the number of data is limited.
2. The proposed method shows that there is relatively strong Pearson correlation between the constructed employment vulnerability index and another related economic indicators. Thus we suggest that the method can be considered to use as an alternative way to survey based approach to monitor the employment vulnerability.

Future Directions

1. Instead of binary classification, we may stratify the employment vulnerability index by it's strength, e.g. high, medium, low risk – however, it depends on the purpose.
2. Enhancement on spatial information, e.g. specific area.
3. Enhancement on more sectors.