Can Media and Text Analytics Provide Insights into Labour Market Conditions in China?

Jeannine Bailliu, Xinfen Han, Mark Kruger, Yu-Hsien Liu, Sri Thanabalasingam
International Department
Bank of Canada staff working papers provide a forum for staff to publish work-in-progress research independently from the Bank’s Governing Council.

This research may support or challenge prevailing policy orthodoxy.

Therefore, the views expressed in this paper are solely those of the authors and may differ from official Bank of Canada views.

No responsibility for them should be attributed to the Bank.
Labour statistics among China’s worst

- The prize for the **dodgiest figures** goes to the labour market
  - Urban unemployment rate is “meaningless” Economist (2008)
  - Wage figures are also “lousy”

- Surveys suggest official rate **underestimates unemployment**
  - Knight and Xue (2007)
  - Wang and Sun (2014)

- Compared to other major countries, China’s official unemployment rate shows **little sensitivity** to changes in output
  - Lam et al. (2015)

- Three relatively high frequency indicators capture formal employment, but not migrant workers
  - Migrant workers could make up 25% of urban employment
    - Wang and Wan (2014)
Our database

- **Chinese language** newspaper database
  - Wisers, a Hong Kong-based company

- We focus on **subset of 90** Chinese newspapers
  - Continuously published over **January 2003 to June 2017**
  - Broad geographic coverage
    - 26 out of 34 regions
    - 77% population

- Building the relevant article pool
  - 8 millions articles from predefined keywords search
  - Downloaded all articles from randomly selected one day per month
    - 266,414 potentially relevant articles
Text mining methodology

1. Training / Testing Data Setup Stage
   - Preprocess articles
   - Bag of words
   - tfidf feature matrix

2. Training Classifier Stage
   - Train SVM
   - Evaluate classifier

3. Machine Classification Stage
   - Preprocess articles
   - Bag of words
   - feature vector
   - Classification

4. Index Construction Stage
   - Full database
   - Labels:
     - Positive 1
     - Negative -1

- Our approach is inspired by Tobback et al. (2016)
  - Use text mining to produce economic policy uncertainty index
Our Labour Market Conditions Index (LMCI)

- SARS & SOE reform
- New Labour Contract Law
- Sichuan earthquake & "Olympic effect"
- Massive layoffs during GFC
- Working age population begins to shrink
- Fiscal stimulus
- Overcapacity cuts
- 5 years plan for poverty alleviation
Text mining methodology

- Why we use machine learning approach?
  - Manual classification **costly**
    - 3 or 4 authors read and classify articles independently
    - Discuss disagreements until consensus reaches
  - Machine learning classification more **consistent**

- Challenges parsing Chinese text
  - In English, unique words are easy to identify since they are separated by spaces
  - Chinese text has no spaces between characters and a character, on its own, may not form a meaningful unit
  - Harbin LTP **natural language processing software**

- Our methodology is **generic** and can be applied to other classification problems
LMCI Validation

- Construct formal models to evaluate LMCI
  - **Wage** Phillips Curve
    - The co-movements between our LMCI and wage growth
  - **McCallum Rule** (1998) with “Chinese characteristics”
    - The PBOC responds in a counter-cyclical fashion to labour market conditions

- Construct two **regional sub-indices**
  - Our results show labour conditions in **coastal regions sensitive** to **export growth**, while in inland regions are not.

- Our study suggests that text analytics can be used to **extract useful labour market information** from Chinese media.