OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

IPS 243
An Alternative Approach for Getting Investment Direction with the Combination of Unstructured and Structured Data

Anggraini Widjanarti, Nursidik Heru Praptono, Mohammad Khoyrul Hidayat
Bank Indonesia

Monday 17 July, 4:00 P.M. – 5:25 P.M.

The views expressed here are those of the authors and do not necessarily reflect the views of Bank Indonesia

OUTLINE

BACKGROUND AND GOALS

BACKGROUND

- **Investment is one of the main indicators to measure state of the economy in Indonesia** (30%, in Indonesia's GDP). However, GDP and Investment are published and available on a quarterly basis with a publication lag of one month.
- Bank Indonesia provides large value payment system, RTGS (Real Time Gross Settlement), that can **generate data related to fund transfers, including investment transactions**.
- Previous research has shown evidence that **investment decision** influenced by sentiment data from newspaper and social media.
- Advancements of technology have opened the **opportunity to explore large dataset of payment data as structured data and news data as unstructured data for monitoring economic activity**.

GOALS

Combining information from unstructured data (news) **and** high frequency **structured data** (large value payment system) **as an alternative approach for investment direction** in Indonesia, by utilizing Big Data Analytics methodology, particularly text mining.

DATA SOURCE

UNSTRUCTURED

- **News Data from 30 prominent newspaper that curated by Bank Indonesia from 2017 – now.**
- **Average total articles are 850 articles/day ≅ 27.000 articles/month**

STRUCTURED

Data Source : BI-RTGS Transactional Data

Transaction Type : Banks' transactions o/b of customer (TTC 100 & 101) ~ MT103

Total Transactions : ≅ 800.000 transactions/months (out of 1,2 mio transactions)

Settlement Time	Transaction Type Code (TTC)	Sender Bank	Receiver Bank	Nominal	Block4
Transaction settlement time (MM/DD/YYYY HH:mm)	**RTGS transaction code** **100 :** Transaction o/b of customer **101 :** Transaction o/b of customer (w/o account) **111 :** Forex Buy/Sell **112 :** Interbank Money Msrket **etc.**	**Sender's Bank SWIFT Code (BIC)** e.g.: CENAIDJA	**Receiver's Bank SWIFT Code (BIC)** e.g.: BMRIIDJA	Transaction amount in Rupiah	Transaction message, sent by Sender Bank into BI-RTGS system

Settlement Time	TTC	Sender Bank	Receiver Bank	Nominal	Block4
5/2/2018 13:35	100	DXXXIDJA	MXXXIDJA	xx,x0,000,000.00	:20:02RE2018043XXXXX#:23B:CRED#:23E:SDVA#:32A:180502IDRxxx0000000,00# :50K:XYZ JAYA,PT#JL. JEND GATOT SUBROTO KAV 51-52#10270 JAKARTA PUSAT#INDONESIA#:52D:BANK DXXX JAKARTA INDONESIA# :53A:/D/521067000990#DXXXIDJA#:57A:/C/520008000990#MXXXIDJA# :59:/121000XXXXXXX#UVW ALIH DAYA, PT# :70:INV. 1804-0167#R/LOCAL# :71A:OUR#:72:/CODTYPTR/100#/CLRC/0670304#:77B:/FEAB/R /PTR/LOCAL

METHODOLOGY

Unstructured Data: News -- *(as the **leading** indicator)*

Text pre-processing → Bag of words → Tfidf feature extraction → Feature vector → Model training & evaluation → model

Filtered, Annotated Data

1 Model Preparation Stage

Full Database

Filtering → Text pre-processing → Bag of words → Tfidf feature extraction → Feature vector → classification → Inferred data → Index construction

Filtered Data

2 Inference on full data & index construction stage

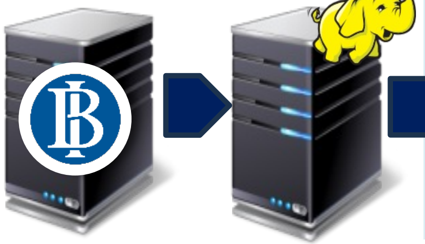
Pos, neg, (sector, foreign/domestic)

3 Validation & analysis stage

Structured Data: RTGS -- *(as the **prompt** indicator)*

1 Data Preparation Stage

BI-RTGS → Hadoop



2 Transaction Data Analysis Stage

Parsing: Extract the field sender, receiver, and transaction description

Entity resolution: Identify the unique entity. Potentially different writing form

Classification: Segmentation/classification of the transactions with text mining.

Aggregation: Summarise the identified transactions

3 Validation & analysis stage

METHODOLOGY UNSTRUCTURED DATA

Model Preparation

- **Filtering:** Filter out non-Indonesian language news**,** tokenize the news into sentences and filter the sentences that contain any keywords that may be related to investment.
- **Data annotation (for model training):** For the model training purpose, we annotate the filtered data. We obtain 1200 positive (+1) and 873 negative or not related (-1). Examples:

menurutnya, ketertarikan grup asal Jepang untuk melakukan investasi di sejumlah proyek energi menunjukkan iklim investasi di indonesia sangat baik. *(**In English:** according to them, the enthusiasm of a group from Japan to invest on some energy projects shows that the investment climate in Indonesia is very good)*	**Positive (+1)**
para investor tidak tertarik lagi melakukan eksplorasi, sehingga produksi minyak nasional akan terus turun. *(**in English**: the inverstors are no longer interested in doing exploration so that the national oil production will continue to decrease)*	**Negative (-1)**

<u>Annotated Data's wordcloud:</u>

Positive (in English): will, able, stock, fund, etc..

Negative (in English): will not, risky, crisis, etc.,

- **Text Preprocessing:** Case folding, remove punctuation and digit, remove stopwords, tokenize into words
- **Feature Extraction:** tfIdf weighting $tfidf_w = tf_w \, * \, \frac{N}{df_w}$

METHODOLOGY UNSTRUCTURED DATA

Inference Model for Sentiment Classification

- Our task is a very **domain specific**. If leveraging anyway any pretrained LLM such as IndoBERT can potentially be misleading. e.g.:

  Tarif pajak yang turun memberikan pengaruh bagi iklim investasi
  *(**in English**: the decrease of tax rate gives impact into the investment climate)*

  Generally, "turun" (decrease) is a sign of negative sentiment. In our case, the decrease of the tax rate is a positive sign of investment sentiment. Thus, we need to do some customizations for our domain.

- **Data annotation can be expensive**! But naturally, we have prior knowledge √ →may be elaborated for the inference
- **Some methods may be limited**: on learning and/or inference only on specific model e.g. (Schapire, 2002) Boosting/Logistic Regression, Lauer & Bloch (2007), Sun Wei (2019) SVM.
- Inspired by Fang & Chen (2011) on **simpler mechanism**: to incorporate **prior knowledge on feature representation** layer, but **we relax to test on some possible models** rather than only one model as theirs' (SVM).

$$g: x \to T_\theta x \text{ so that } f(x) = f(g(x))$$

"Add a token to promote if any word exists in dictionary"

- Berpikir bahwa investasi sector manufaktur merupakan peluang **menarik** + **positiveInvestment**

  (Think that investment in manufacture is an interesting chance)

- Investor **khawatir** akibat krisis ini + **negativeInvestment**

  (the investors are wory for this crisis)

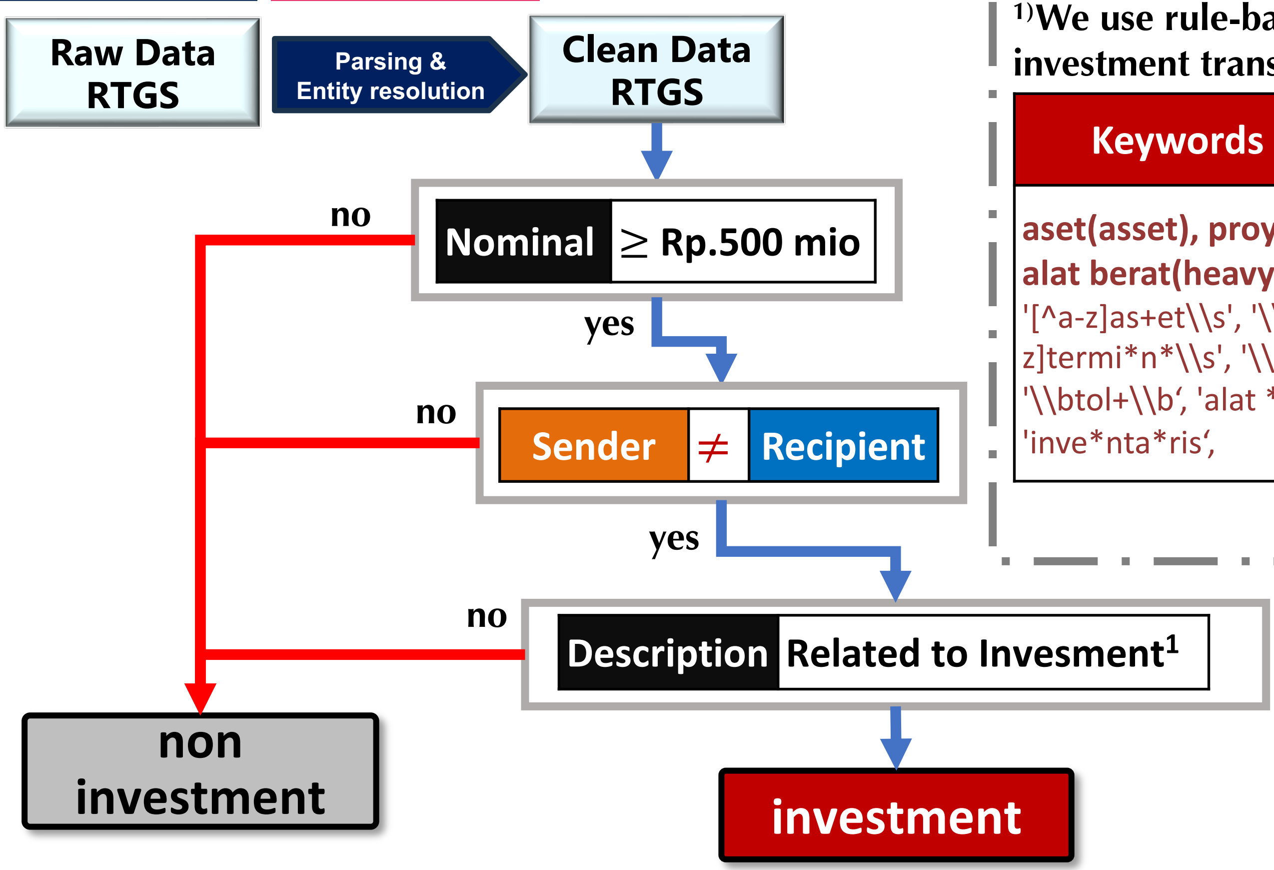
Experimental Result

Model (f)	without Prior Knowledge		with Prior Knowledge	
	F1	Error	F1	Error
SVM	80,62	24,24	**83,60**	**20,67**
Logistic Regression	79,59	24,07	82,53	20,74
Decision Tree	74,30	30,03	77,71	25,91
XGBoost	81,10	23,37	82,69	21,20

**note: with SMOTE (Chawla et al 2002)*

METHODOLOGY STRUCTURED DATA

Classification of RTGS Investment Transactions



[1]We use rule-based (keyword) approach for classifying investment transaction

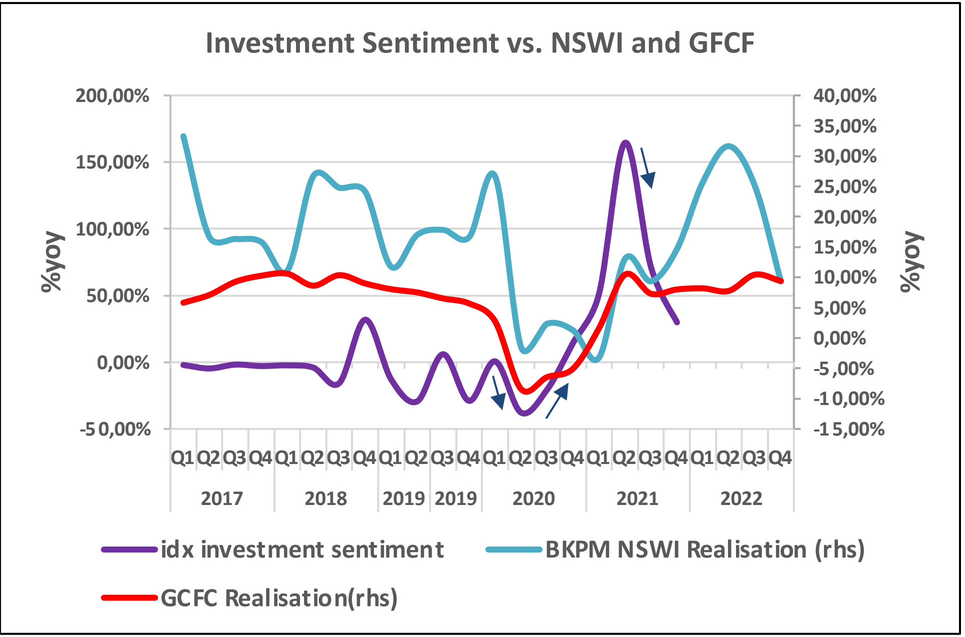
Keywords that are related to investment
aset(asset), proyek (project), tol(toll), termin(term), alat berat(heavy equipment), investaris(inventory): '[^a-z]as+et\\s', '\\bas+et\\b', '\\sas+et[^a-z]', 'proyek', '[^a-z]termi*n*\\s', '\\btermi*n*\\b', '\\stermi*n*[^a-z]', '\\btol+\\b', 'alat *be*ra*t','pe*ra*la*ta*n *(tek|tech|ind)', 'inve*nta*ris',

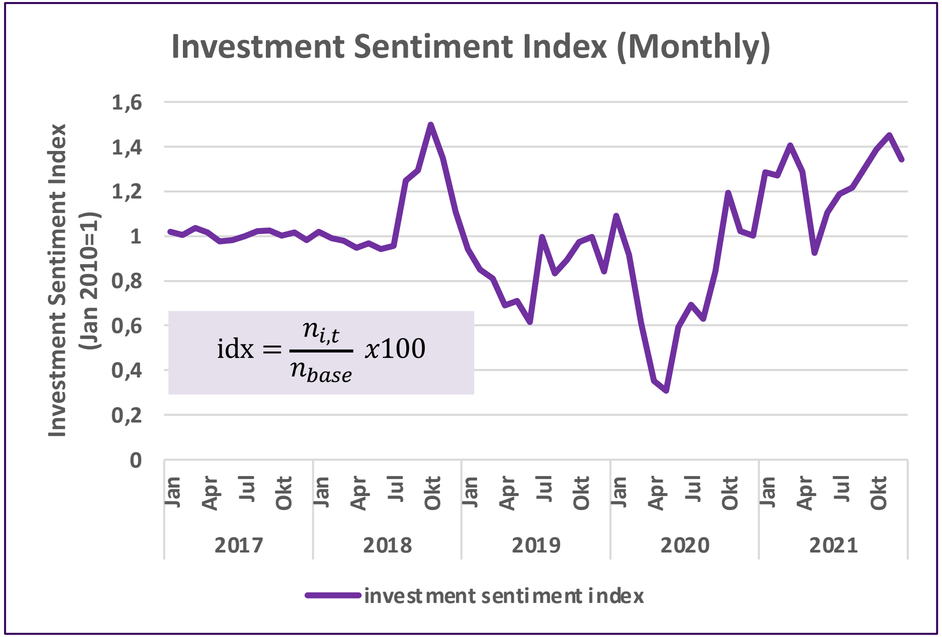
Keywords that are not related to investment
transfer dana, pindah dana, pindah buku (fund transfer/overbook): 'tra*n*s*fe*r*\\s*dana', 'pi*nda*h\\s*dana', '\\bpb\\b', 'pi*nda*h\\s*buku', '\\bpinbuk\\b'

RESULT AND ANALYSIS

Investment Sentiment Index from News for *leading indicator*

The validation results show that by the lag of 4 quarters, there are positive correlations between investment sentiment index vs. NSWI ($\rho\_Pearson$=0,77) and GFCF ($\rho\_Pearson$=0,33).

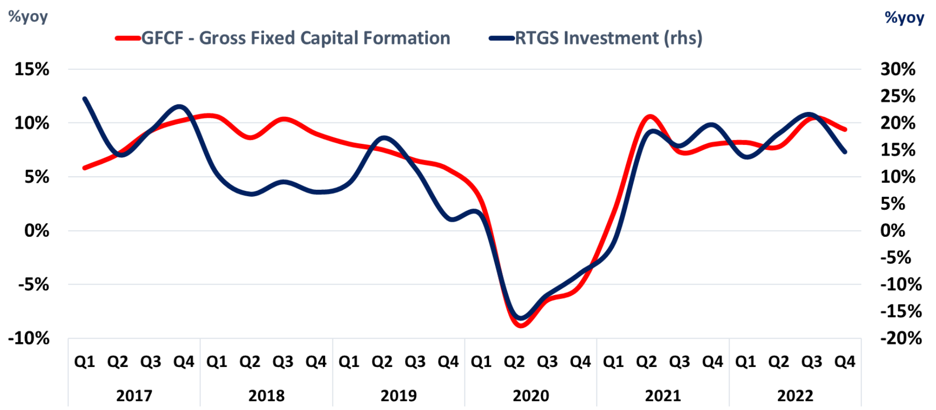




RESULT AND ANALYSIS

RTGS Investment Transactions for prompt indicator

The validation results show a high correlation between the two indicators since quarter 1-2017, including during the Covid-19 pandemic and subsequent period.



Indicators	Correlation of RTGS Investment Growth Rate with GFCF	
	Q1-2017 s.d. Q4-2022	Q1-2020 s.d. Q4-2022
Total of Gross Fixed Capital Formation	85,4%	97,6%
GCFC – Construction	83,6%	95,0%
GCFC – non Construction	78,6%	93,8%

CONCLUSION AND FUTURE WORKS

Conclusion

1. We have **proposed and alternative approached** in utilizing unstructured and structured data to construct indicators that helped to track investment direction in Indonesia.
2. Our leading investment indicator, **using text mining methodology from news through machine learning method,** can help to see the investment direction trends in Indonesia (up to 1 year). **Based on the evaluation metrices score on the model, the average F1-score of the model is 83,6 % with error rate is 20,7%**
3. Our prompt investment indicator can be **generated more quickly from payment system data using the proposed text mining methodology** compared to the GFCF indicators in GDP publications. The validation results demonstrate a high correlation between our investment indicator from the payment system and the GFCF indicators, indicating that the payment system indicator can be served as a reliable proxy for prompt investment indicators.

Future Works

1. **Developing method for sectoral investment indicators**, especially to track main sectors in Indonesia.
2. **Model accuracy improvement** e.g. further model tune up
3. **Constructing a nowcasting model for investment indicator,** involves incorporating indicators of news, payment system data, and other macroeconomic variables through the use of machine learning algorithms or econometric techniques.

OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

THANK YOU.