Irving Fisher Committee on
Central Bank Statistics

BIS

11th Biennial IFC Conference on “Post-pandemic landscape for central bank statistics”

BIS Basel, 25-26 August 2022

Measuring payment system policy credibility using machine learning[1]

Muhammad Abdul Jabbar, Okiriza Wibisono and Alvin Andhika Zulen,
Bank Indonesia

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

Measuring Payment System Policy Credibility Using Machine Learning

Okiriza Wibisono[1], Muhammad Abdul Jabbar[2], Alvin Andhika Zulen[3]

Abstract

In this digital age, payment system plays a crucial role in ensuring the smooth functioning of economic and financial activities. There have been many developments in the payment system landscape, such as digital / wallet-based payments, QR for payment, open APIs, and fast payment systems. For some consumer segments, all these progress can be daunting so as to hinder adoption. On the other hand, some consumers may use these payment facilities without being aware of how best to protect their funds and data. In order to realize the benefits while managing the risks of payments in the digital era, central banks need to understand how the public perceives the payment system ecosystem and its related policies. In our previous research, we have developed machine learning methodology to measure central bank monetary policy credibility. In this paper, we apply a similar machine learning methodology using news data to measure Bank Indonesia's payment system policy credibility. In our case, the index measures public sentiment on 5 aspects of payment system policy: infrastructure, industry conduct, regulation / policy formulation, entry policy, and supervision. The index is aimed at helping central banks in formulating better payment system policy, e.g. by informing better policy communication to the public and as feedback for policy formulation process.

Keywords: payment system, policy credibility, natural language processing

JEL classification: E42, C88

The views and results expressed here are those of the authors and do not necessarily represent Bank Indonesia.

[1] Department of Statistics, Bank Indonesia. email: okiriza_w@bi.go.id

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

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

Contents

1. Background

In this digital age, payment system plays a crucial role in ensuring the smooth functioning of economic and financial activities. For example, during 2021, SMS/mobile and internet banking in Indonesia reached almost 8 billion transactions in volume, growing 57% year on year (Bank Indonesia Payment Systems and Financial Market Infrastructure Statistics – Table 7). There have been many developments in the payment system landscape in Indonesia and globally, such as digital/wallet-based payments, QR for payment (Beck et al., 2022), open APIs (BIS-BCBS, 2019), and fast payment systems (BIS-CPMI, 2021). The COVID-19 pandemic end the ensuing mobility and contact restrictions further accelerate noncash payments worldwide (BIS, 2021). These progresses are all welcome, since digital financial inclusion in payments is shown to boost economic growth by as much as 2.2 percentage points (Khera, 2021). In this regard, Bank Indonesia is well-prepared to navigate various developments in payments and related sectors, having launched our Indonesia Payment System Blueprint 2025 back in 2019.

In order to realize the benefits while managing the risks of payments in the digital era, there is a need to build awareness and ascertain consumers that adopting innovative payments technology can be safe and beneficial for them. For some consumer segments, recent innovations in payments can be daunting so as to hinder adoption. Various studies have shown that consumers' adoption of new technology, such as mobile payments, are driven by the consumers' perceived risk of said technology. This can include perceived financial risk, privacy risk, performance risk, psychological risk, and time risk. These risks in turn can be affected by perceived technological uncertainty, information asymmetry, regulatory uncertainty, and service intangibility (Yang, 2015).

Other line of research has shown that mobile payments adoption are also driven by the way consumers obtain information. (Suoranta, 2004) find that for adopting new technology, experienced users are more compelled by interpersonal communication by service providers. Non-users and less experienced users, on the other hand, receive information more from mass media, which is related to this paper's topic. Thus, understanding how the public perceives their payment system policy and the payment ecosystem more broadly can help central banks in formulating payment system policy.

For this purpose, Bank Indonesia used to regularly conduct survey to external stakeholders to measure policy credibility, including payment system policy credibility. The Policy Credibility Survey is based on 6 aspects of credibility: formulation, independence, communication, accountability, coordination, and effectiveness.

In practice, the survey method (in general) has several weaknesses:

1. Survey fatigue: respondents experiencing burnout if surveyed repeatedly, so that surveys cannot be carried out too often (especially since the pool of economists or other stakeholders as respondents can be quite limited);
2. Desirability bias: respondents giving a response that is favorable for the surveyor, which is the central bank, hence the results are less objective;
3. Recency bias: respondents generally providing responses based on recent policies and/or events, hence the survey results are very dependent on the execution time; and
4. Survey cost & time.

Based on these considerations, we develop a payment system policy credibility measurement (indexes) by utilizing Big Data Analytics. The indexes are constructed from text mining of public perceptions toward payment system and payment system policy credibility that are reported in news media. This paper draws largely from the methodology in (Wibisono, 2022). We apply a similar machine learning methodology using news data to measure Bank Indonesia's payment system policy credibility. The resulting index measures public sentiment on 5 aspects of payment system policy: infrastructure, industry conduct, regulation / policy formulation, entry (policy), and supervision (these aspects are aligned with various payment system functions within related department in Bank Indonesia). The index is aimed at helping central banks in formulating better payment system policy, e.g. by informing better policy communication to the public and as feedback for policy formulation process.

The paper is organized as follows. In section 2, we provide literature reviews on Bank Indonesia's Policy Credibility Survey and text mining for policy analysis. In section 3, we discuss the data and methodology. In section 4, we provide a summary of the results and evaluation of the model. In section 5, we conclude the paper and offer some thoughts for future works.

2. Literature Review

2.1 Bank Indonesia's Policy Credibility Survey

From 2013 to 2018, Bank Indonesia conducted Bank Indonesia's Policy Credibility Survey, a semi-annual survey to measure policy credibility for all 3 sectors of policy: monetary, macroprudential, and payment system. The survey was aimed to provide a measure for policy credibility that is objective, accurate, reflecting broad view of stakeholders (including general public), and available timely. The survey was used to determine the effectiveness of policy communication as well as feedback for formulating future policy communication strategies.

The target respondent of the survey was approximately 1,000 respondents in 20 major cities in Indonesia, consisting of government personnel, bankers, industry players, academics, and general public. The survey measured 6 aspects of policy credibility, from which our indexes are derived:

1. Formulation: whether our policies are formulated carefully according to their objectives
2. Independence: whether we formulate our policies independently, without intervention from any party
3. Communication: whether our policies are well-communicated to the public
4. Accountability: whether our policies are well accounted for
5. Coordination: whether we always coordinate well with the government
6. Effectiveness: whether our policies are effective in achieving their objectives

While most of the aspects above are based on the literature on central bank credibility, especially monetary policy credibility, they are perhaps less relevant for measuring payment system policy credibility, especially independence and coordination aspects. Thus, in this paper we use different credibility aspects, which we describe in section 3.

2.2 Text mining for policy analysis

Text data have been widely used for research in economics and finance. Nowadays, text mining algorithms are growing rapidly along with the adoption of big data and machine learning. These algorithms can automatically "read" and “extract” relevant information from texts, such as person's name, topics, and sentiment. Compared to manual approach, text mining allows us to make use of much larger text data faster, including news, social media, and press releases. Applied to news media, text mining has the potential to complement survey indicators by extracting and quantifying public’s opinions and sentiments contained in the news.

As an example related to central banks, Sahminan (2008) identified keywords that reflect a tight, neutral, or loose monetary policy inclination in the press release statement of Bank Indonesia over the period from January 2004 to December 2007. (Tobback et al., 2017) developed the Hawkish-Dovish (HD) index that measures media’s perception of ECB communications using two methods: semantic orientation (SO) and support vector machine (SVM). These are based on co-occurrences of strings and machine learning classification algorithm.

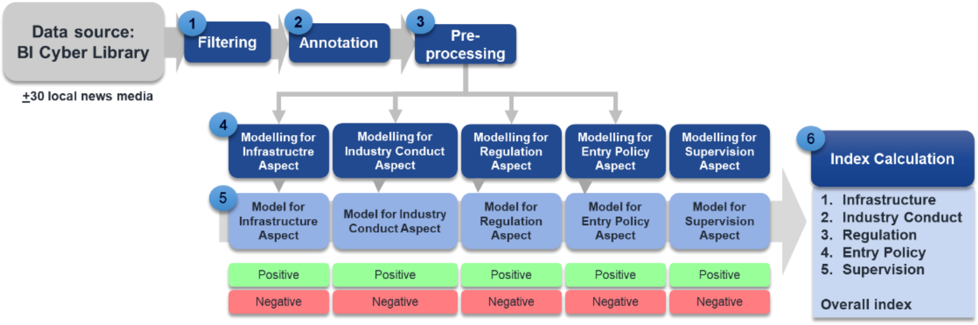
Recently, we developed text mining methodology for measuring monetary policy credibility from news (Wibisono, 2022). We use news that are relevant to monetary policy; specifically, sentences that contain any of the related keywords: “inflation”, “monetary”, “exchange rate”, “current account”, “policy rate”, “BI Rate", "BI 7-Day Reverse Repo Rate", and their variations in writing. Using methodology similar to what we describe in the next section, the unstructured news sentences are passed into machine learning model and the outputs are aggregated into policy credibility indexes.

3. Methodology

The overall methodology for constructing the policy credibility indexes starting from source data is depicted in Figure 1.

Overall methodology

Figure 1


Data source:
BI Cyber Library
±30 local news media
1
Filtering
2
Annotation
3
Pre-processing
4
Modelling for Infrastructre Aspect
Modelling for Industry Conduct Aspect
Modelling for Regulation Aspect
Modelling for Entry Policy Aspect
Modelling for Supervision Aspect
5
Model for Infrastructure Aspect
Model for Industry Conduct Aspect
Model for Regulation Aspect
Model for Entry Policy Aspect
Model for Supervision Aspect
Positive
Negative
6
Index Calculation
1. Infrastructure
2. Industry Conduct
3. Regulation
4. Entry Policy
5. Supervision
Overall index

3.1. Data

Source: News data serves as the main input for constructing the policy credibility indexes. We use news data from Bank Indonesia's Cyber Library, which is a curated internal repository of news articles related to economic and financial topics. There are more than 30 local news media, with an average of about 850 articles daily, although the number of news media and news articles can vary from month-to-month. The news data are available on a daily basis since 1999, but the news data that we use in this paper span from January 2013. The news are in Bahasa Indonesia (Indonesian language).

Filtering: We filter out news that are not relevant for constructing the payment system policy credibility index. Specifically, we only keep news *sentences* that contain any of the keywords related to payment systems. In total there are more than 100 keywords used, examples of which are: "payment system", "BI-RTGS", "BI-FAST" (our fast payment system), "internet banking", "mobile banking", "e-money", "payment service providers", "card payments", "credit cards", "debit cards", "merchant discount rate", "transfer fee", "QRIS" (our standard for QR payment), and "Open API". Furthermore, the sentence or its previous/next sentence must mention "BI" or "Bank Indonesia".

After an initial run of our model development, we expanded the list of keywords to include those that pertain to payment system regulation, entry policy, and supervision. These are important functions within Bank Indonesia and thus we want to be able to monitor news about these three topics as well. Example keywords are "consumer protection", "payment service licensing", "supervisory technology", "fraud supervision", and "cyber security".

3.2. Policy Credibility Index

3.2.1. Annotation

A random sample of the filtered news sentences are manually annotated to construct training data for "teaching" machine learning classification models. Each sentence is labelled with 5 information, representing public's perception on the payment system credibility aspects. The possible labels for each aspect are positive, negative, or irrelevant, indicating whether the perception/sentiment contained in the sentence are positive, negative, or unrelated to the aspect.

The 5 aspects and their short descriptions are:

1. Payment system infrastructure: policy, developments, and conduct (e.g. reliability, safety, efficiency) of payment system infrastructures, both those that are operated by Bank Indonesia such as BI-RTGS and BI-FAST, or by the industry such as National Payment Gateway (GPN)
2. Payment system conduct by industry: conduct of payments services by the industry, how Bank Indonesia's payment system policies are implemented by the industry
3. Payment system regulation: whether Bank Indonesia's payment system policies are well-formulated and their effectiveness in achieving their intended objectives

4. Payment system entry policy: effectiveness and efficiency of payment system entry and licensing activities
5. Payment system supervision: Bank Indonesia's supervision of the payments industry, e.g. related to payment service and payment infrastructure providers, consumer protection

Annotation is done by the authors and subject matter experts on payment system policy within Bank Indonesia. Prior to annotation, we write out the guidelines on how to annotate the news sentences including specific examples, so that the result is more consistent across annotators. Each sentence is annotated by 2-3 annotators to minimize bias.

A total of 8,556 sentences were annotated. Example annotated sentences for each aspect are provided in Appendix A.

We initially targeted for 5,000 annotated sentences, but additional set of sentences were annotated since we found that there were few sentences that have negative sentiment label, and also few sentences that are relevant to entry policy and supervision aspects. Furthermore, the label distributions after additional annotation are still heavily imbalanced towards positive label, so we resort to machine learning modeling technique that deals with imbalanced data as described in section 3.2.3.

Distribution of annotated sentences Table 1

Credibility aspect	Positive	Negative	Irrelevant
Infrastructure	561 (6.6%)	24 (0.3%)	7,971 (93.1% of total)
Industry conduct	1,745 (20.4%)	200 (2.3%)	6,611 (77.3% of total)
Regulation	1,423 (16.6%)	80 (1.0%)	7,053 (82.4% of total)
Entry policy	105 (1.2%)	27 (0.3%)	8,424 (98.5% of total)
Supervision	179 (2.1%)	42 (0.5%)	8,335 (97.4% of total)
TOTAL	4,013 (9.4%)	373 (0.09%)	41,789 labels (97.7%)

In general, there are very few sentences with negative label. For payment system entry policy and supervision aspects, almost all sentences are irrelevant.

3.2.2. Data preprocessing

Each filtered sentence (not necessarily annotated) as described in section 3.1 is transformed from textual format into tabular-numeric so that it can be processed by machine learning algorithms, following the steps below:

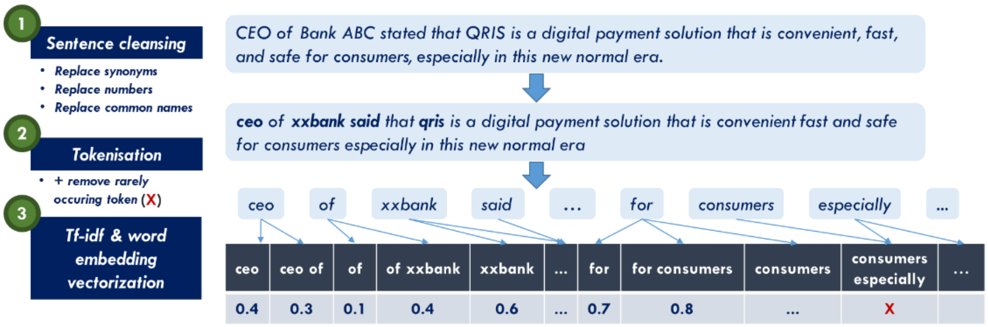
1. Sentence cleansing: lowercasing, replacing synonyms, abbreviations, numbers, and common names in the sentence;
2. Tokenization: splitting sentence into words/tokens, removing rarely occurring terms;

3. N-gram tf-idf vectorization: creating n consecutive words (n-gram) with term frequency-inverse document frequency as features (Juan, 2003);
4. Word embedding vectorization: creating vector as numeric representation for each sentence through the average of its words' word embedding vector (Mikolov et al., 2013);

An example is shown in Figure 2 below.

Example data (sentence) preprocessing steps

Figure 2


1
Sentence cleansing
• Replace synonyms
• Replace numbers
• Replace common names
2
Tokenisation
• + remove rarely occuring token (X)
3
Tf-idf & word embedding vectorization
CEO of Bank ABC stated that QRIS is a digital payment solution that is convenient, fast, and safe for consumers, especially in this new normal era.
ceo of xxbank said that qris is a digital payment solution that is convenient fast and safe for consumers especially in this new normal era
ceo
of
xxbank
said
…
for
consumers
especially
…
ceo
ceo of
of
of xxbank
xxbank
…
for
for consumers
consumers
consumers especially
…
0.4
0.3
0.1
0.4
0.6
…
0.7
0.8
…
X

3.2.3. Model training

From the preprocessed and annotated sentences, we train machine learning models to classify each sentence into one of possible labels (positive/negative/irrelevant). So in total we have 5 sets of models representing the 5 aspects.

Handling imbalanced labels: As can be seen from the annotation results (Table 1), the class distributions are quite imbalanced, with most sentences in the news being irrelevant (even after filtering with payment system-related keywords). For relevant sentences, the distributions are also heavily imbalanced towards positive class. Considering these imbalances, we carry out the classification in 2 stages for each credibility aspect:

1. Classifying whether the filtered sentence contains sentiment about payment system (either positive or negative, vs. irrelevant); and
2. For relevant sentences (those that contain sentiment), classifying the sentiment in the sentence (positive vs. negative).

In addition, we also applied synthetic-minority oversampling technique (SMOTE) to remedy the label imbalances (Chawla, 2002).

Cross validation: We use 5-fold cross validation: the sentence features generated in section 3.2.2 and resampled with SMOTE are divided into training and validation sets, with approximately 80% in training set (the exact percentages differ across credibility aspects to accommodate the different label distributions). We repeat this 5 times, so we have 5 different training and validation sets, to get a more robust measure of model accuracy.

Model training: The sentences in the training sets are input into machine learning algorithms. We experimented with various machine learning algorithms: logistic regression, k-nearest neighbors, support vector machine (SVM), naïve bayes, decision tree, random forest, XGBoost, and deep learning – long short-term memory (LSTM). To obtain the best accuracy, each algorithm is constructed with various hyperparameter settings.

In total we will have 10 machine learning models, since we have 2 machine learning stages (classifying relevance and classifying sentiment) for each of the 5 aspects.

Evaluation: The resulting models are given the sentences in the validation sets, to measure their accuracy on unseen data. We use macro-average F1 score across labels as evaluation metric:

$$F1 = average(F1_l), l \in \{positive, negative, irrelevant\}$$

$$F1_l = precision_l \times recall_l$$

$$precision_l = \frac{\#true\ positive_l}{\#predicted\ positive_l}$$

$$recall_l = \frac{\#true\ positive_l}{\#annotated\ positive_l}$$

3.2.4. Text classification and index calculation

Aspect index: Having obtained the classification models for each credibility aspect, we apply the models to the whole (filtered) news sentences in Cyber Library, including those that were not annotated. For each time period (quarterly or yearly), we tabulate the number of sentences classified as positive or negative for each credibility aspect.

The index for each aspect is calculated as the net balance of the number of positive and negative sentences in each time period.

$$index_{aspect,t} = \frac{\#positive_{aspect,t} - \#negative_{aspect,t}}{\#positive_{aspect,t} + \#negative_{aspect,t}}$$

The indexes have the characteristics of a net balance index, as below:

1. Range of index: [-100%,100%].
2. The index will be close to 100% if there are more news sentences with positive sentiment on the policy credibility aspect. Conversely, the index will be close to -100% if there are more news with negative sentiment.
3. Positive index means more news sentences with positive sentiment on the policy credibility aspect, compared to negative sentences. Conversely, negative index means more news with negative sentiment than positive ones.
4. Zero index means equal number of news sentences with positive sentiment and negative sentiment on the policy credibility aspect.
5. If $index_{t1}$ > $index_{t2}$ then the proportion of news with positive sentiment on the policy credibility aspect is greater in $t1$ than in $t2$. Conversely, if $index_{t1}$ < $index_{t2}$ then the proportion of news with negative sentiment on the policy credibility aspect is greater in $t1$ than in $t2$.

Overall index: The 5 indexes are weighted-averaged to obtain the overall payment system policy credibility index, for each period (quarterly or annually). The weights are based on the aspects' number of sentences. In addition, we apply a threshold on the number of relevant sentences on each aspect: there must be at least 4 sentences in a quarter. The threshold value is chosen by median number of sentences in each aspect over the whole data. Aspects with 3 or less sentences in a quarter will be excluded from the overall index.

The full formula for the overall index is below, where $\mathbb{1}$ denotes the thresholding (whether there are more relevant sentences than the threshold).

$$index_t = \sum_{aspect} \mathbb{1}_{aspect,t} \times (\#positive_{aspect,t} + \#negative_{aspect,t}) \times index_{aspect,t}$$

Besides on historical data, we also calculate the index calculation for ongoing periods, without the need for more manual annotation as the sentence classification models have been trained.

4. Result and Discussion

4.1 Machine learning model evaluation

Below is the best result for each credibility aspect, averaged across the validation sets.

Machine learning model evaluation results

Table 2

Credibility aspect	Best model combination	A: Relevance classification F1	B: Sentiment classification F1	End-to-end F1 (A*B)
Infrastructure	logistic regression & logistic regression	75%	86%	64.5%
Industry conduct	logistic regression & XGBoost	72%	78%	56.2%
Regulation	logistic regression & logistic regression	73%	74%	54.0%
Entry policy	logistic regression & random forest	75%	88%	66.0%
Supervision	XGBoost & decision tree	68%	89%	60.5%
Overall (average)	-	72.6%	83.0%	60.2%

Several observations that can be made from the above results are:

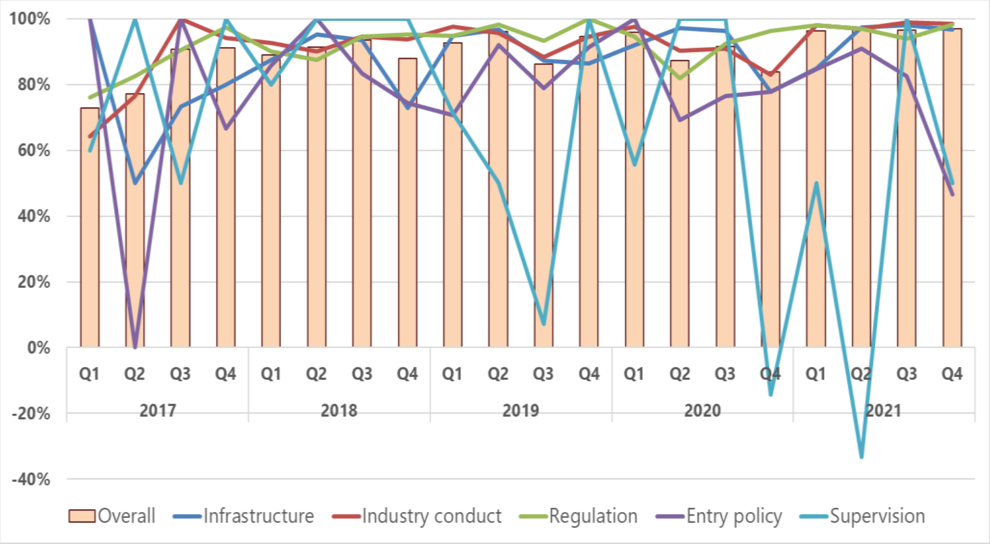
- Logistic regression algorithm is the most accurate in majority of cases. This suggests that the relationship between sentence features and credibility labels is mostly linear, or that more (annotated) data is needed to be able to extract more accuracy from using nonlinear algorithms.
- Classifying sentiment (positive vs. negative) is relatively easier than classifying whether the sentence contains sentiment in the first place (83.0% vs. 72.6% averaged macro-F1). Thus, further model development can be focused on improving relevance classification.
- Payment system industry conduct and regulation aspects have lowest end-to-end F1 (due to lowest sentiment F1). This is somewhat unexpected, since these aspects have the largest share of negative sentences (less imbalance), although the numbers are still quite small. We posit that this result may be due to the evaluation metric used, and needs to be further inspected.
- End-to-end F1 (60.2% average) is acceptable, but may warrant further improvement to ensure more robustness of the resulting credibility indexes.

4.2 Index results

The resulting indexes are presented in Figure 3. The lines represent each credibility aspect, and the bar is the overall (threshold-weighted-averaged) index.

Payment system policy credibility index

Figure 3



Referring to the figure, we comment on some of the indexes' movements:

- Most indexes are always positive, which means that there are more positive sentences about payment system than negative sentences in the news. The overall index is also increasing over time (86.1% in 2017 to 96.7% in 2021; annual numbers not shown in figure).
- Only supervision index ever reaches negative value. However, we note that supervision index has high volatility, and to less extent, entry policy index as well. This reflects the small number of sentences relevant to these indexes as described in Table 1, even after we applied SMOTE algorithm during model training. This is also one of our considerations for thresholding and weighting the aspect indexes when calculating the overall index.
- The indexes have average pairwise correlation of 13.4% (30.7% if excluding the volatile supervision index), although industry conduct and regulation has 75% correlation. On a high level, the low average correlation suggests that the aspect indexes contain different information that can be relevant for payment system policy analysis.

Example issues classified correctly by the model in 2022 (index numbers not shown in Figure 3) are in Table 3 below.

Example issues identified in 2022 by aspect

Table 3

Credibility aspect	Example issues
Infrastructure	Positive news about our newly implemented fast payment system, BI-FAST, its features and advantages
Industry conduct	Positive news about wider acceptance and use of QRIS (QR Indonesia Standard), and about the increase in QRIS transaction limit
Regulation	Positive news about Bank Indonesia's payment system policy, which is aimed at accelerating payment system digitalization to further integrate the digital economy and support national economic recovery
Entry policy	Positive news about payment system providers licensing by Bank Indonesia
Supervision	N/A (index cannot be calculated since the number of relevant sentences is less than threshold)

5. Conclusion & Future Works

5.1. Conclusion

We develop a methodology for measuring Bank Indonesia's payment system policy credibility by utilizing news articles data and machine learning-based technique. From the out-of-sample evaluation results, we achieve an average F1-score of 60.2%. The methodology is largely drawn from our previous research on monetary policy credibility index from news (Wibisono, 2022).

The resulting policy credibility index shows a positive trend, which means that news about payment systems in recent years is more positive. The machine learning models also seem to be able to capture relevant payment system developments from news, such as BI-FAST implementation and wider adoption of QRIS. On the other hand, some of the aspect indexes, i.e. entry policy and supervision, are highly volatile, which mostly likely is due to the small number of sentences relevant to these aspects.

5.2. Future works

Some possible research directions include:

- **Reidentification of credibility aspects:** In our current methodology, we have 5 policy credibility aspects: payment system infrastructure, industry conduct, regulation, entry policy, and supervision. As described before, these are aligned with the different functions performed by related department in Bank Indonesia. However, there are at least 2 issues with the current aspect grouping: (1) entry policy and supervision aspects have very few relevant sentences, and (2) industry conduct and regulation aspects show high correlation. Further assessment may be needed to find the best set of aspects that are relevant to payment system policy-making, have low pairwise index correlations, and whose sentences are relatively easy to classify.
- **Model improvement:** As noted before, end-to-end average F1-score of the machine learning models is 60.2%. On average, this means that the models have 40% probability to miss relevant sentences or misclassify the sentiment (positive vs. negative) of relevant sentences. For more robust policy credibility indexes, there may be a need to improve the accuracy of the models. Several options to improve accuracy are: collect more annotated data (e.g. through additional

keywords) and use nonlinear algorithms in conjunction with the larger data. We can also try to divide the keywords into specific keywords for each aspect to reduce irrelevant sentences, since we observe that the number of irrelevant sentences (Table 1) can be as few as 1.5% per aspect.

- **Econometric analysis:** This paper focuses on developing the policy credibility indexes, from source data, sentence filtering, annotation, preprocessing, machine learning modeling & evaluation, and finally calculating the indexes. Our proposition is that the indexes, at the highest level, provide a measure of how well payment systems and payment system policy is covered in the news. This may impact the tendency for consumers to use/adopt innovative payments services, as well as impact their trust on the central bank's payment system policy. It will be interesting to analyze the significance of the payment system policy credibility indexes, e.g. as explanatory variable in the framework of mobile payments adoption presented in (Yang, et al., 2015) or analysis of drivers of digital financial inclusion presented in (Khera, et al., 2021).
- **Updating of topics/keywords**: Compared to other policy areas of central bank, e.g. monetary and macroprudential policy, it can be argued that developments in payment systems are faster, technological or otherwise. Thus it is a challenge to keep up with relevant issues about payments in the news, let alone quantify them as positive or negative. Methods in textual trend detection can be considered as a solution for this (Kontosthatis, et al., 2004).

References

Bank for International Settlements (2021) commentaries. Covid-19 accelerated the digitalisation of payments.

Bank for International Settlements – Basel Committee on Banking Supervision (2019). Report on open banking and application programming interfaces (APIs).

Bank for International Settlements – Committee on Payments and Market Infrastructures (2021). Developments in retail fast payments and implications for RTGS systems. *CPMI Papers* no. 201.

Beck, T., et al. (2022). Big techs, QR code payments and financial inclusion. *BIS Working Papers* no. 1011.

Chawla, N.V., et al. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, vol. 16, pp. 321-357.

Juan, R. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first institutional conference on machine learning*, vol. 242, no. 1, pp. 29-48.

Khera, P., et al. (2021). Is digital financial inclusion unlocking growth? *IMF Working Paper*, no. 21/167.

Kontostathis, A., et al. (2004). A Survey of Emerging Trend Detection in Textual Data Mining. *Survey of Text Mining*, pp. 185-224, Springer.

Mikolov, T., et al. (2013). Efficient estimation of word representations in vector space. arXiv preprint no. 1301.3781.

Sahminan. (2008). Effectiveness of Monetary Policy Communication in Indonesia and Thailand. *BIS Working Paper* No. 262.

Suoranta, M. & Mattila, M. (2004). Mobile banking and consumer behavior: new insights into the diffusion pattern. *Journal of financial services marketing*, vol. 8, no. 4, pp. 354-366.

Tobback, E., Nardelli, S., & Martens, D. (2017). Between Hawks and Doves: Measuring Central Bank Communication. *ECB Working Paper Series* No. 2085.

Wibisono, O., et al. (2022). Machine learning for measuring central bank policy credibility and communication from news. *IFC workshop on "Data science in central banking" - Part 2: Data Science in Central Banking: Applications and tools*.

Yang, Y., et al. (2015). Understanding perceived risks in mobile payment acceptance. *Journal of international management & data systems*, vol. 115, no. 2, pp. 253-269.

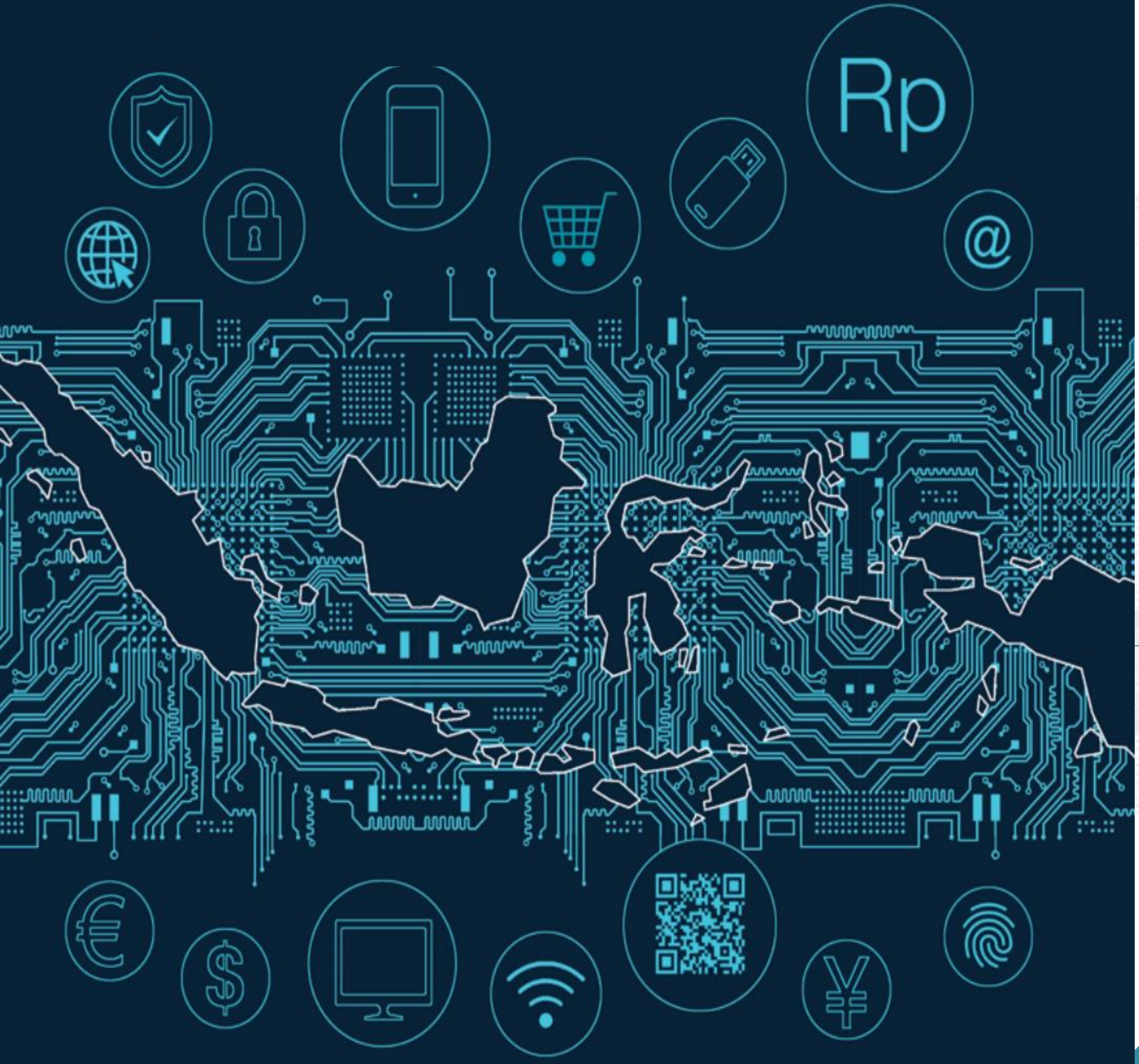
Appendix A: Example annotated sentencs

Example annotated sentences from news

Table A.1

No.	Sentence*	Credibility aspect	Label
1.	*Bank Indonesia strengthens payment system infrastructures and promotes noncash payments.*	**Infrastructure**	***Positive***
2.	*Bank Indonesia detected that on Tuesday there have been network disturbance in their BI-RTGS, BI National Clearing System (SKNBI) and BI Scripless Securities Settlement System (BI-SSSS).*	**Infrastructure**	***Negative***
3.	*Since the standardization of electronic payments using QRIS, about 4 million MSMEs have adopted QRIS [for accepting payments].*	**Industry conduct**	***Positive***
4.	*QRIS hampers payment system efficiency since there is requirement to become interoperable and interconnected.*	**Industry conduct**	***Negative***
5.	*Bank Indonesia's payment system policy is aimed at accelerating payment system digitalization to further integrate the digital economy and support national economic recovery.*	**Regulation**	***Positive***
6.	*Regulation on fintech companies by Bank Indonesia and by Financial Services Authority have not been able to provide legal certainty for consumers.*	**Regulation**	***Negative***
7.	*We [Bank Indonesia] are reviewing [payment system] licensing system so that it can be more efficient.*	**Entry policy**	***Positive***
8.	*There are complaints from the industry that the fintech licensing system [by Bank Indonesia] is complex and costly in terms of time required.*	**Entry policy**	***Negative***
9.	*Bank Indonesia closely supervises payment service providers in order to ensure consumer protection in the payment system and digital economy.*	**Supervision**	***Positive***
10.	*Bank Indonesia must recover consumer's loss in this incident of payment system failure.*	**Supervision**	***Negative***

*) Translated by the authors from Bahasa Indonesia to English. Notes in [square bracket] are added for clarity.



BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

Measuring Payment System Policy Credibility Using Machine Learning

Okiriza Wibisono, Muhammad Abdul Jabbar,
Alvin Andhika Zulen

11th IFC Biennial Conference
25-26 August 2022

The views and results expressed here are those of the authors and do not necessarily represent Bank Indonesia.

Background and Goals

***Payment system plays a crucial role** in ensuring the smooth functioning of economic and financial activities. More so in this digital age.*

***Public's perception on the payment system** ecosystem may impact their adoption of new developments in payments.*

Previous approach to measure public perception on our payment system policy credibility: semiannual survey to stakeholders (e.g. economists, academics, government, general public).

*This research: Utilizing **Big Data Analytics** – text mining to gather public perception regarding payment system ecosystem and its related policies.*

Methodology largely based on our previous use case on measuring monetary policy credibility (Wibisono, 2022).

Data & Overview of Methodology

News articles

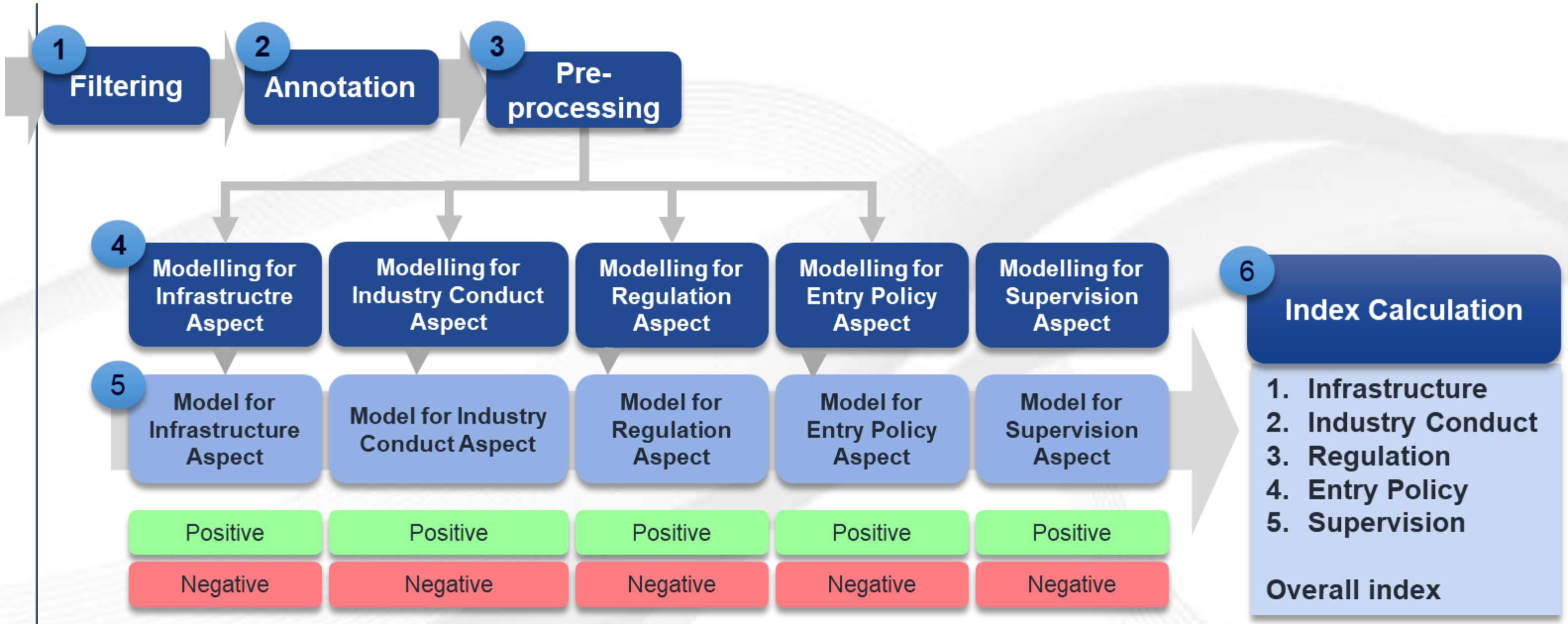
Source: Cyber Library (internal repository of curated economic and financial news)

~30 domestic news (in Bahasa Indonesia)
~850 articles daily

Whole corpus: since Jan 1999
Training data: Jan 2013 – Sep 2021

Example keywords for filtering the news:

- payment system
- BI-FAST
- BI-RTGS
- SKNBI
- internet banking
- mobile banking
- e-money
- payment service providers
- card payments
- credit cards
- debit cards
- transfer fee
- EDC
- Open API
- QRIS
- consumer protection
- payment service licensing
- supervisory technology
- fraud supervision
- cyber security



1 Filtering

2 Annotation

3 Pre-processing

4 Modelling for Infrastructre Aspect | Modelling for Industry Conduct Aspect | Modelling for Regulation Aspect | Modelling for Entry Policy Aspect | Modelling for Supervision Aspect

5 Model for Infrastructure Aspect | Model for Industry Conduct Aspect | Model for Regulation Aspect | Model for Entry Policy Aspect | Model for Supervision Aspect

Model for Infrastructure Aspect	Model for Industry Conduct Aspect	Model for Regulation Aspect	Model for Entry Policy Aspect	Model for Supervision Aspect
Positive	Positive	Positive	Positive	Positive
Negative	Negative	Negative	Negative	Negative

6 Index Calculation

1. Infrastructure
2. Industry Conduct
3. Regulation
4. Entry Policy
5. Supervision

Overall index

5 Credibility Aspects

Payment system infrastructure: policy, developments, and conduct (e.g. reliability, safety, efficiency) of payment system infrastructures, both those that are operated by BI or by the industry

Payment system conduct: conduct of payments services by the industry, how BI's payment system policies are implemented by the industry

Payment system regulation: whether BI's payment system policies are well-formulated and effective in achieving their intended objectives.

Payment system entry policy: effectiveness and efficiency of payment system entry and licensing activities

Payment system supervision: BI's supervision of the payments industry, e.g. related to payment service and payment infrastructure providers, consumer protection

Methodology – Index Construction

1. Annotation

A sample of filtered sentences are annotated as training data for ML classification models.

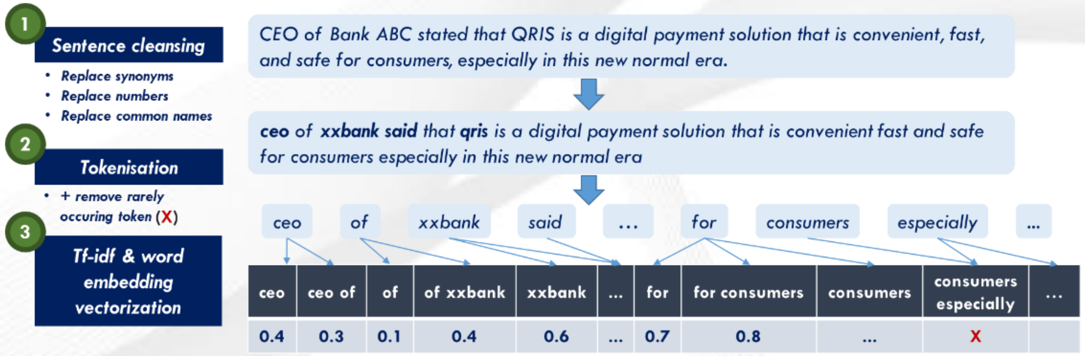
- ➢ Annotated by authors and domain experts within BI.
- ➢ Guidelines incl. examples
- ➢ 2-3 annotators per sentence
- ➢ Most sentence turns out irrelevant for the aspect

Aspect	Positive	Negative	Irrelevant
Infrastructure	561 (6.6%)	24 (0.3%)	7,971 (93.1%)
Industry conduct	1,745 (20.4%)	200 (2.3%)	6,611 (77.3%)
Regulation	1,423 (16.6%)	80 (1.0%)	7,053 (82.4%)
Entry policy	105 (1.2%)	27 (0.3%)	8,424 (98.5%)
Supervision	179 (2.1%)	42 (0.5%)	8,335 (97.4%)
TOTAL	4,013 (9.4%)	373 (0.09%)	41,789 (97.7%)

2. Data preprocessing

Each sentence is transformed from text into tabular-numeric format for training ML models.

- ➢ Sentence cleansing
- ➢ Tokenization
- ➢ Remove sparse terms
- ➢ N-gram vectorization
- ➢ Word embedding



CEO of Bank ABC stated that QRIS is a digital payment solution that is convenient, fast, and safe for consumers, especially in this new normal era.

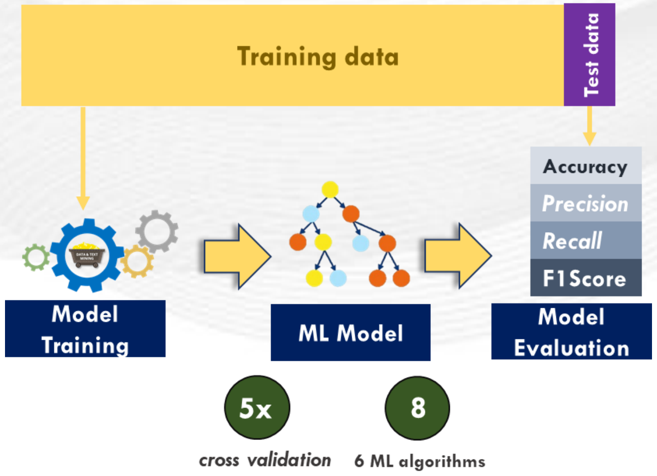
***ceo** of **xxbank said** that **qris** is a digital payment solution that is convenient fast and safe for consumers especially in this new normal era*

ceo	ceo of	of	of xxbank	xxbank	...	for	for consumers	consumers	consumers especially	...
0.4	0.3	0.1	0.4	0.6	...	0.7	0.8	...	X	

3. Model training

ML model is trained for classifying sentences into pos/neg labels, for each aspect.

- ➢ **5-fold CV**, 80-20 train-validation split
- ➢ **SMOTE** to handle imbalanced labels
- ➢ **2-step classification**: relevant vs irrelevant, positive vs negative
- ➢ Best average **macro F1**: 60.2%
- ➢ Best algorithm: **logistic regression**



4. Index calculation

- The ML models are applied to all sentences, to construct monthly indexes.
- The overall index is a weighted average of the 5 component indexes based on number of sentences.
- Any component with 3 or less sentences in a quarter is excluded.

$$index_{aspect\ k,t} = \frac{\#positive_{k,t} - \#negative_{k,t}}{\#positive_{k,t} + \#negative_{k,t}}$$

$$index_t = \sum_{aspect} \mathbb{1}_{aspect,t} \times (\#positive_{aspect,t} + \#negative_{aspect,t}) \times index_{aspect,t}$$

Example annotated sentences (only relevant sentences)

No	Sentence	Credibility Aspect	Label
1	*Bank Indonesia strengthens payment system infrastructures and promotes noncash payments.*	Infrastructure	*Positive*
2	*Bank Indonesia detected that on Tuesday there have been network disturbance in their BI-RTGS, BI National Clearing System (SKNBI) and BI Scripless Securities Settlement System (BI-SSSS).*	Infrastructure	*Negative*
3	*Since the standardization of electronic payments using QRIS, about 4 million MSMEs have adopted QRIS [for accepting payments].*	Industry conduct	*Positive*
4	*QRIS hampers payment system efficiency since there is requirement to become interoperable and interconnected.*	Industry conduct	*Negative*
5	*Bank Indonesia's payment system policy is aimed at accelerating payment system digitalization to further integrate the digital economy and support national economic recovery.*	Regulation	*Positive*
6	*Regulation on fintech companies by Bank Indonesia and by Financial Services Authority have not been able to provide legal certainty for consumers.*	Regulation	*Negative*
7	*We [Bank Indonesia] are reviewing [payment system] licensing system so that it can be more efficient.*	Entry policy	*Positive*
8	*There are complaints from the industry that the fintech licensing system [by Bank Indonesia] is complex and costly in terms of time required.*	Entry policy	*Negative*
9	*Bank Indonesia closely supervises payment service providers in order to ensure consumer protection in the payment system and digital economy.*	Supervision	*Positive*
10	*Bank Indonesia must recover consumer's loss in this incident of payment system failure.*	Supervision	*Negative*

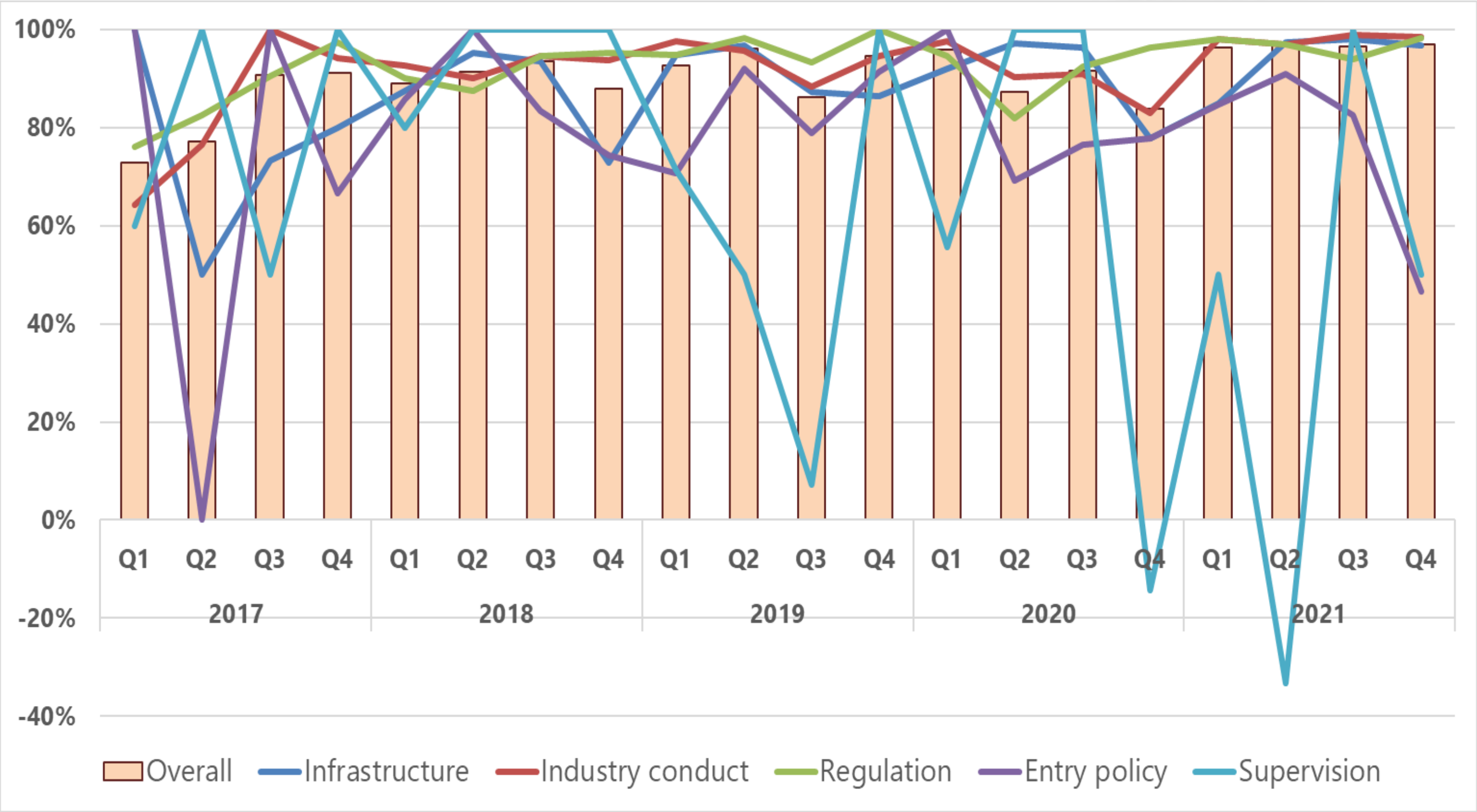
Result – ML model

Credibility aspect	Best model relevance	Best model sentiment	A: Relevance classification F1	B: Sentiment classification F1	End-to-end F1 (A*B)
Infrastructure	logistic regression	logistic regression	75%	86%	64.5%
Industry conduct	logistic regression	XGBoost	72%	78%	56.2%
Regulation	logistic regression	logistic regression	73%	74%	54.0%
Entry policy	logistic regression	random forest	75%	88%	66.0%
Supervision	XGBoost	decision tree	68%	89%	60.5%
Overall (average)	-	-	**72.6%**	**83.0%**	**60.2%**

Some observation about the results:

- ❑ **Logistic regression** algorithm is the most accurate in majority of cases.
- ❑ Classifying **sentiment** (positive vs. negative) is relatively **easier** than classifying whether the sentence contains sentiment in the first place (83.0% vs. 72.6% averaged macro-F1).
- ❑ Payment system **industry conduct and regulation aspects** have lowest end-to-end F1 (due to lowest sentiment F1). This is somewhat unexpected, since these aspects have the largest share of negative sentences (less imbalance).
- ❑ **End-to-end F1** (60.2% average) is **acceptable**, but may warrant further improvement to ensure more robustness of the resulting indexes.

Result – Indexes



Some observation about the indexes:

- ❑ **Most** of the indexes are always **positive**, which means that there are more positive sentences about payment system than negative sentences in the news.
- ❑ The overall index is also **increasing** over time (86.1% in 2017 to 96.7% in 2021; annual numbers not shown in figure).
- ❑ **Only supervision** index ever reaches **negative** value. However, we note that supervision index has **high volatility**, and to less extent, entry policy index as well.
- ❑ The indexes have average pairwise **correlation** of 13.4% (30.7% if excluding the volatile supervision index), although industry conduct and regulation has 75% correlation.
- ❑ **Example** recent trends captured by the models: positive news about our newly implemented payment system (BI-FAST), positive news about wider use of QRIS.

Conclusion and Future Works

Conclusion

1 Developed a machine learning methodology for measuring public's perception of payment system policy credibility by utilizing news data.

2 The resulting index shows positive trend: according to the models, news about payment systems in Indonesia in recent years is more positive. But supervision index is highly volatile due to small number of relevant sentences.

3 The models seem to be able to capture relevant developments, such as BI-FAST implementation and wider adoption of QRIS.

Future Works

1 Reidentification/redefinition of the credibility aspects.

2 Model accuracy improvement e.g. by annotating more data, or using specific keywords for each aspect.

3 Econometric analysis (econometric effect of the indexes on macro indicators).

4 Developing method for automatic updating of keywords.