11th Biennial IFC Conference on "Post-pandemic landscape for central bank statistics"
BIS Basel, 25-26 August 2022

Sentiment analysis of user's reviews on non-bank payment service apps,[1]

Muhammad Hafiruddin, Mohammad Khoyrul Hidayat,
Arinda Dwi Okfantia and Nursidik Heru Praptono,
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.

Sentiment Analysis of User's Reviews on Non-Bank Payment Service Apps

Nursidik Heru Praptono[1,2], Arinda Dwi Okfantia[1], Mohammad Khoyrul Hidayat[1], Muhammad Hafiruddin[1]

Abstract

The digital payment activities on non-bank payment service environment have shown to be increasing, especially during the COVID-19 pandemic. Any systemic risks on monetary and payment system stability affected by mobile apps quality should be anticipated as early as possible. We present an inference model to analyse the user's review sentiments on mobile apps quality on some aspects, address some related inference problems, and construct an index based on the inferred sentiments. Having some promising results, we suggest that our approach can be used by policy makers in order to timely monitor the performance of non-bank payment service providers.

Keywords: Probabilistic Inference, Machine Learning, Limited Training Data, Non-Formal Text, Text Mining, Sentiment Analysis, User Reviews, Mobile Apps, Non-Bank Payment Service Provider, Monetary and Payment System Stability.

JEL classification: C44, E42

-The views expressed in this work belong to the authors and do not necessarily reflect the institution-

[1] Statistics Department, Bank Indonesia.

[2] nursidik_hp@bi.go.id/pra.heru@gmail.com (corresponding email)

Contents

1. Background

More banking and payment activities are currently being conducted into mobile application platform than the traditional ones. Such services can legally be provided by the authorised entities, either bank or non-bank institutions. A number of nonbank payment service providers are increasingly running these services. In Indonesia for example, by the end of 2021, there have been at least 41 non-bank payment service providers (PSP) who take a part in this area. Through mobile application platforms that they provide, an amount of services related to the payment activities are offered including for example e-wallet and e-money. The digital economy is thus accelerated through this landscape.

The growing amount of non-bank PSP customers using such services has also been found to be increasing recently, moreover since the pandemic of COVID-19. By the beginning of year 2022, at Google Play (for Android users), more than 100 million installs non-bank PSPs were reported. In addition to that, at AppStore (for Apple Iphone users) although the number of installs of this top non-bank PSPs is not explicitly reported, the number of reviews reached more than 800 thousand. Considering the whole 41 providers, the non-bank PSP's have taken a part and infiltrated on nearly the whole of about 270 million people of the Indonesian population. This amount describes that the payment services provided by non-bank PSP are non-trivial thing and have a significant role in promoting and accelerating economic growth. They however require authorisation, supervision and monitoring from Bank Indonesia under the Payment System Policy.

This massive implementation of payment services provided by non-bank PSP is not without challenges. Many risks can occur if there is any problem on running such services. Of this phenomena we identify that there are some perspectives that need to concern about based on the entities involved (customers, non-bank PSPs, and policy makers):

1. Customers however need the reliable mobile application provided, it is intuitive that they would either oversee the review before install and/or give any reviews on the mobile apps' performance.
2. Non-Bank PSPs have to secure their reputational issue so that their business strategy can run properly as planned. Any problem related to the reliability of their mobile apps would matter, since mobile apps take the closest proxy of their representation to the customers. Thus they need to be able to provide reliable payment services mobile apps for the customers.
3. Policy maker concerns about monetary stability and payment system security & efficiency. It is important for policy maker to oversee the performances of non-bank PSP's since any problem on running payment services through mobile apps can potentially impact the stability in the macroeconomic perspective. This is to enable the policy maker to consider an early decision to the non-bank PSP's in advance.

Measuring the quality of mobile apps of payment services therefore becomes important as it represents the quality of non-bank PSP. However, performing such measurement itself is not a straightforward way to do. A proxy for this measurement is by leveraging user review's sentiment on the quality of mobile apps on some aspects. Analysing sentiment is not without any problem. Through this paper, we

propose a methodology in order to analyse the user's sentiment of such mobile apps. Generally, the contribution of our paper can be listed as the following:

1. We conduct sentiment analysis approach as the proxy for mobile payment apps measurement of nonbank PSP, as thus a proxy to oversee the condition of the non-bank payment system in the macroeconomic perspective.
2. We perform the comparison of several inference models and address some issues related to their performance due to the possibility of limited training data access.
3. We construct series of index based on the inferred sentiments.

The organisation of this paper is as follows: Section 2 provides literature review related to the assessment of mobile apps that may be related to our domain problem, Section 3 describes our methodology to infer the sentiment and to construct the sentiment's index. Section 4 presents our experimental results and analysis. Finally, we conclude our works and discuss future direction in Section 5.

2. Literature Review

The importance of quality assessment of internet based proxy for banking or financial related activities can initially be described in the work by Rod et al (2009). In their work SEM PLS methodology was leveraged in order to analyse the questionnaire they gathered. Their work showed that customer service quality, online information system quality, and banking service product have positive impact to the customer satisfaction.

Following their works, a similar research goal had also been conducted by Ganguli et al (2011). A generic service quality dimension assessment of the internet banking was conducted through an exploratory factor analysis. They found four important aspects: customer services, technology security and information quality that affect the customer satisfaction, while technology convenience, and easiness and realibility that affect the customer loyalty.

The banking services are however expanding into mobile banking due to the freedom of time and place. Therefore mobile banking then obviously becomes a promising measurement source for proxy. For the banking sustainability reason, it is thus important to retain bank's customer's loyalty. Thakur (2013) investigated some factors that could have positive impact on customer's loyalty to the bank. The work gives analytical result that mobile interface usability and service had positive impact on customer satisfaction. Arcand et al (2017) later, utilised SEM and found that security and practicity impact on trust, while enjoyment and sociality impact the commitment and satisfaction. The related works were also performed by Sagib and Zapan (2014), Rahman et al (2017), and Khan et al (2021), on the measurement of m-banking service quality.

Another work on more general domain related to the mobile apps quality assessment method can be seen in the work by Vu et al (2015). They proposed a methodology assessing user's opinion on mobile apps that can be the concern for app's developer. An amount of keywords related to the was found, and the work gives the result that the technology is one of the important factors to concern in order to improve user's experience and satisfaction.

Related to the bank's reputation, Bach et al (2020) investigated some aspects that determine the relationship between mobile banking and bank reputation. Their work leverage 500 clients on a number of local banks in Croatia. The work suggest that safety, simplicity, and service variability of mobile banking application gives positive impact to the bank reputation.

It is thus important to analyse the user's sentiment on mobile payment apps. In addition to the work of Vu et al (2015), there have also been another related works related to the customer's sentiment. Fang and Zhang (2015), Singla et al (2017) conducted the on sentiment analysis on review data of Amazon product utilising some NLP and machine learning techniques. These works however share the same property in the perspective of measurements of a product quality: utilising the customer's opinion

Recently, and more related to our case, the work of Leem & Eum (2021) utilised text mining in order to analyse the sentiment of mobile banking service quality. They leveraged user's review obtained from application store. There were 5 aspects/dimension they investigate which are security/privacy, practicity, design/aesthetics, sociality and enjoyment. The analysis reported that the customers mostly complained about four topics which are process, interaction, customer convenience, and technology and function.

Considering the above conducted researches, we then consider some aspects related to the mobile apps applied to non-bank PSP. The case of bank PSP and non-bank PSP however shares the same property: utilising mobile apps in financial and payment system area. In our work, we propose an alternative, quick and timely methods as the proxy for qualitative measurements of non-bank PSP using text mining. Of those aspects mentioned in the literature, we deterministically define 5 aspects related to the mobile apps in order to demonstrate the sentiment analysis:

- **Security Aspect**: Represents the security of mobile apps. including e.g. some issue related to the problem on security aspects including log in/log out, registration, security of e-money.
- **Feature Aspect**: Represents the variability and functionality of the features offered by non-bank PSP's.
- **Design Aspect**: Represents how well the design related to the intuitiveness and user experiences.
- **Customer Service Aspect**: Represents how responsible, reliable, and the effectiveness of the customer services.
- **Technology Aspect**: Represents how advanced the technology the non-bank PSP's has, how well the application is from the technological issues.

3. Methodology

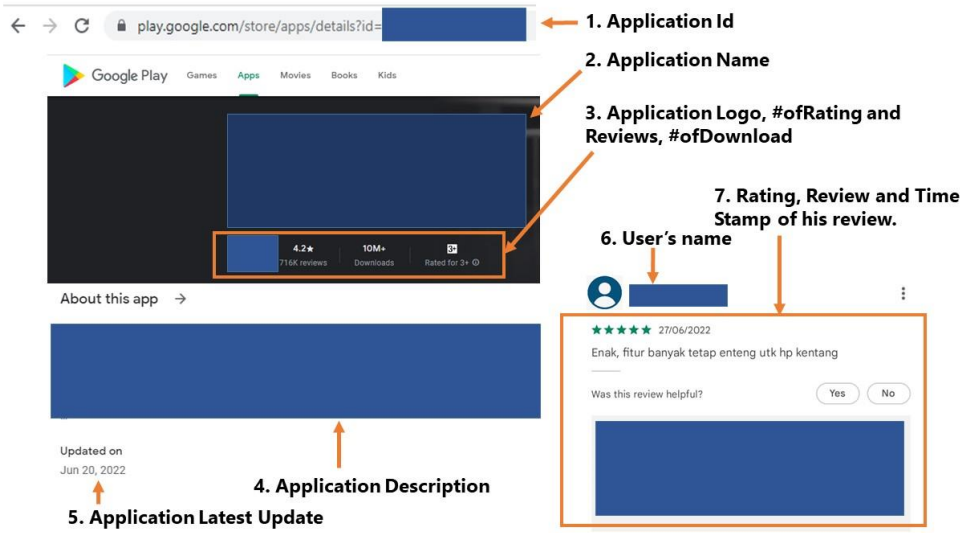
3.1. Data Collection and Annotation

Data Collection

In order to conduct the experiment, first we collect the data from Goole Play Store and Apple Appstore by scrapping methods, with Python programming language. The structure of information available can be seen in Fig 1 below.

Available Information on the Store

Figure 1



* case study: Google Play Store, for Android Users

In general, the information that is available at the Google Play Store as on Fig 1 above consists of application id (appId), application name, application logo, number of rating reviews, number of download, application's description, it's latest update, user's name, rating, review and time stamp of his/her review. The information available on Apple Appstore is however relatively similar. Of those information, for each of the appId we then extract the user's review based on the username, name of user, and time stamp.

From about 41 nonbank-PSP's total population, we obtained 35 mobile apps of nonbank-PSP available on the store, therefore the number of nonbank-PSP investigated though our work is 35. The total of the review scraped on Google Play Review and Apple AppStore for the period of December 2011 up to December 2021 in this experiment is 4.637.980 reviews, with language in Bahasa Indonesia.

Data Annotation

In order to develop and to validate our inference model, we annotate a number of samples. We sample 1270 datapoints (i.e. reviews) across the players over the monthly period. To this data we flag both aspects and sentiment's magnitude. Here, a review may contain more than one aspect. For example, a review may discuss a bad feature of mobile app, but good at the user interface design. The distribution of this annotated dataset can be seen as on Table 1 below.

The sentiment category formulation in our case is slightly different compared to other sentiment analysis case study in general. For each aspect, we put the positive or not related aspect to be in the same category, while negative ones are categorised as negative,. This is due to the purpose, that the policy maker may prefer to see only the negative sentiment. The challenge in our case is then the limited amount of training data as we are dealing with new specific domain problem in the specific language: mobile apps' user's reviews in Bahasa Indonesia.

Annotated Dataset Table 1

	Negative	Positive or Not Related	Total
Security Aspect	449	821	1270
Feature Aspect	420	850	1270
Design Aspect	306	964	1270
Customer Service Aspect	390	880	1270
Technology Aspect	427	843	1270

[1] Annotation is performed by 3 annotators.

3.2. Data Preprocessing

Text Normalisation.

The review obtained are of less formal lexical form. It is because the users are free to write anything expression as the review. Lexical normalisation is then an important stage so that the further feature extraction will become more optimum. We construct a dictionary and leverage deterministic regular expression in order to normalise the unnormalised lexical form. Any digit, punctuation and emoticon are removed. In addition to that, any review with length less than 25 characters are eliminated as after our inspection, most of them are meaningless/noise. Some examples of unnormalised vs. normalised text can be seen as in Table 2.

Unnormalised vs. Normalised Text Table 2

Unnormalised Text (raw text)	Normalised Text (cleaned text)
KAK SAYA SUDAH MELAKUKAN ISI SALDO VIA KLIK XX. SALDO SAYA DI XX TERPOTONG TAPI KOK TIDAK MASUK DI SALDO yy.. TERUS UANGNYA NYANGKUT DIMANA???? mohon bantuannya	kak saya sudah melakukan isi saldo via klik xx saldo saya xx terpotong tapi kok tidak masuk saldo yy terus uangnya nyangkut dimana mohon bantuannya *(in English: I have bought the credits phone using xx and the balance is debited, but no yy points added I got, where is my money? please help*)
kecewa banget keamanan kurang, akun XX saya bisa disadap orang jadi kena penipuan ah anjir kesellllll 1juta hilang	kecewa banget keamanan kurang akun xx saya bisa disadap orang jadi kena penipuan anjir kesell juta hilang *(in English: very disappointed for bad security my xx account was tapped so I got scammed damn I lost milions)*
Aplikasi yg Bagus.. Tampilannya menarik..	aplikasi yang bagus tampilannya menarik (*in English: nice application, yet interesting user interface*)

Feature Extraction.

We further apply tfidf bag of words weighting in order to extract the textual feature. This can be expressed as in the Eq 1 following.

$$tfidf_{w,d} = tf_{w,d}\ x\ \log\frac{N}{df_w} \quad (1)$$

each feature word w is converted to numerical representation according to the number of occurrence of words w in document d that is $tf_{w,d}$ weighted by the log inverse of document frequency i.e. the frequency of document containing words w, that is df_w. Here N is number of all reviews in the collection.

3.3 Inference Models

Rule Based Models (Deterministic Approach)

The first and relatively the most straightforward inference model is to leverage rule based model. We provide predefined keywords that both belongs to aspects, and also sentiments. Some sample of list of keywords can be seen in Table 3 following.

Some keywords example and their translation in English — Table 3

Keywords Category	List of Keywords Example	Translation (English)	Notes
Security Aspect	*log in, log out, penipuan, otp, verifikasi, username, password…*	log in/out, scamming, otp, verification, username, password	In implementation, regular expression (regex) is utilised to reassure that the lexical/wrting variety is captured.
Feature Aspect	*Fitur, bayar, simpan, akses, servis, menu, saldo, emoney, pembayaran…*	Feature, pay, save, access, service, menu, balance, emoney, payment..	
Design Aspect	*tampilan, desain, smooth, scroll, klik..*	(user) interface, design, scroll, click...	
Customer Service Aspect	*aduan, complain, keluhan, tanggapan..*	complain, rant, (customer service's) response..	
Technology Aspect	*memory, lemot, lelet, error, crash,..*	memory, slow, error, crash	
General Negative Setiment	*menyebalkan, menyusahkan, sampah, konyol, lemot, error, crash, jelek…*	annoying, bothering, rubbish, awful, slow, error, bad	
General Positive Sentiment	*bagus, baik, mantap..*	good, nice, excellent..	
Negation	*Tidak, belum, kagak..*	not, not yet..	

In order to infer the aspect and sentiment, we leverage those keywords into the simple rule as shown in Eq 2. Basically, the rule r tries to find aspect and its sentiment based on the lexical occurance given the keywords on a review.

$$r(S) = \exists(\mathrm{kw}_{\mathrm{aspect}})\ \mathrm{in}\ S \bigwedge \exists\ (\mathrm{kw}_{\mathrm{negative\_sentiment}})\ \mathrm{in}\ S \quad (2)$$

Here S is review text and kw is keyword. This rule will return 1 if there is any negative sentiment on the particular aspect. Any negative sentiment that is followed by negation keywords will turn positive, and thus will give 1 value for $r(S)$. However, after our manual inspection, this kind combination is relatively rare, as usually user

states single token opinion sentiment. Although the knowledge representation can be expressed explicitly, this model costs on the pattern complexity: it has to be able to do a generalisation well to accurately extract the information.

Models Built from Data

In order to perform the sentiment using models that rely on the data, we then utilise machine learning. We demonstrate the models which are: Decision Tree with gini index as splitting criteria, SVM with radial basis function (RBF) kernel, Logistic Regression. In our case, each aspect's sentiment is analysed with different models each other. Therefore, we have in total 15 machine learning models to assess (3 models for each aspect's sentiment analysis, and we have 5 aspects). These models are trained and evaluated onto the annotated data at the Table 1. We use 5-fold cross validation scenario in order to evaluate those models. Furthermore, onto the model from data that achieve high result, we also perform experiment on elaborating prior knowledge as discussed on the Result and Discussion Section.

Imbalanced Data Handling

As we have seen on Table 1 that on *per aspect's perspective*, the distribution between negative and non-negative classes (either positive or not related) are relatively imbalance. We then leverage the oversampling technique for the minority class (in this case negative class) with Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al, 2002) to enable the data synthesis. We perform such approach in the feature space, before the parameter estimation of those machine learning models.

Evaluation of the Inference Models

Our problem can be seen as a binary problem, with negative sentiment is our interest magnitude. To evaluate the performance, we use F1 scores as formulated as follows:

$$\text{F1 Score} = \frac{2\,(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

where

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

3.4. Index Formulation

After the full data are completely inferred by applying the best model trained from the annotated data, the next stage is to construct an index for each aspect. A review on each aspect should have either 0 (positive/neutral) or 1 (negative). The index is formulated as shown in Eq 3 below:

$$\text{idx}_{o,t} = \frac{\sum_i 1_{f(x_{0,i,t})=1}}{\text{N}_{o,t}} x \log \text{N}_{0,t} \tag{3}$$

Generally speaking, the index is about the proportion of negative sentiment of a NB-PSP o within time t over the total of its review at the time t that is, $(N_{o,t})$.

In our case study, we consider prefer negative sentiment to positive sentiment. Negative sentiment is rather more useful to the policy makers as they would oversee

any problem so that they could consider further decision, to prevent any subsequent systemic risk on monetary and payment system.

The log term in Eq 3 is to weight the magnitude of the proportion. In other words, it is considered to quantitatively represent how important a NB-PSP is within the share. Intuitively, a number of negative sentiment on the top player should be take care of compared to the player whose small portion of market share.

4. Result and Discussion

4.1. Result on the Inference Models

The experimental result can be seen in Table 4. To note that we also experimented the inference scenario without SMOTE beforehand (that is let the imbalanced data as it is when training the model). However, the result is not better than when the SMOTE is utilised. Therefore, we only show the model when the SMOTE is performed during the training phase. We performed the experiment with 5-fold cross validation.

Performance matrix of models from data (F1 in %) Table 4

	Security Aspect	Feature Aspect	Design Aspect	Customer Service Aspect	Technology Aspect
Rule Based	56,49	56,14	55,43	53,78	55,85
SVM	76,14	74,67	66,67	66,55	*65,77*
Decision Tree	72,00	61,90	63,37	59,36	63,69
Logistic Regression	*78,17*	*74,80*	*69,32*	*79,07*	65,61

According to the performance evaluation, the Logistic Regression model gives highest F1 score in general, at least in 4 aspects. In the Technology aspect however SVM gives the higher result than the Logistic Regression although the score differences is relatively small. The rule based on the other hand gives the lowest performance (less than 60% in all aspects).

Having the performance result as shown in Table 4 above, it can be seen that in general Logistic Regression takes over the other inference models. We then performed further experiment in the next part on how we incorporate the prior knowledge into our Logistic Regression.

Incorporating Prior Knowledge

The models from data as implemented previously rely on the data. We in this case attempt to incorporate our prior knowledge into the model for two reasons. First usually it is hard to access the labelled data to train the model properly i.e. annotation is a costly process. Second, we have prior knowledge related to the information we want to find. In our case we have some keywords and rules as our prior knowledge.

We then demonstrate on how to incorporate the prior knowledge. In general, Logistic Regression can be expressed as Eq 4 follows:

$$p(y = 1|x) = \sigma\big(f(x)\big) = \frac{1}{1 + e^{-f(x)}} \tag{4}$$

where $f(x) = w\phi(x)$ represents our linear function between parameter w and basis function $\phi(x)$. Here $x \in \mathbb{R}^d$ is feature vector representation (our input), and $y \in \{0,1\}$ is the target (our label). The objective function is to minimise negative log likelihood as formulated in Eq 5:

$$J = \sum_i \ln(1 + e^{(-(2y_i - 1)f(x_i))}) \tag{5}$$

We recapture the idea of the work by Schapire et al, 2002 on how to incorporate prior knowledge into the model. First we introduce a distribution that quantifies our prior belief from rule based model, that is:

$$\pi = p(y = 1|x) = \begin{cases} 0.9, & \text{if } r(S_x) = 1 \text{ (True)} \\ 0.1, & \text{otherwise} \end{cases} \tag{6}$$

The objective function as defined on Eq 5 is then modified, becomes:

$$J = \sum_i \ln\big(1 + e^{(-(2y_i - 1)f(x_i))}\big) + \underbrace{\eta D_{KL}(\pi(x_i)||\sigma(f(x_i)))}_{\text{control on prior information}} \tag{7}$$

The second part of the addition in Eq 7 above intuitively represents the control of our prior knowledge. It is quantified by KL Divergence between the prior information π and the information obtained from training data $f(x)$. The variable η controls the importance of the prior information. In our experiment, we leave its value into 1 as in this work it is not our main focus. We then estimate the model's parameter by minimising cost function in Eq 7 by applying BFGS approximation algorithm.

The obtained parameter is then used in the decision function. First, we transform our prior probability into $h_0(x)$ as follows:

$$h_0(x) = \sigma^{-1}(\pi(x)) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) \tag{8}$$

Subsequently, the $h_0(x)$ is blended to the final function: $f^*(x) = f(x) + h_0(x)$. The decision function as expressed in Eq 4 thus becomes $p(y = 1|x) = \sigma(f^*(x))$.

The experimental results of this scenario can be seen as in Table 5 below.

Performance on Logistic Regression with Prior Knowledge (F1 in %) — Table 5

	# of Training Data	Logistic Regression	Prior Knowledge (Rule Based Model)	Logistic Regression with Prior Knowledge
Security Aspect	100	54,11	56,49	55,00
	200	66,67	56,49	67,86
	500	72,64	56,49	75,20
	800	75,76	56,49	78,82
	1016	78,17	56,49	**81,36**
Feature Aspect	100	50,39	56,14	57,65
	200	51,81	56,14	74,90
	500	62,50	56,14	77,67
	800	64,15	56,14	77,87

	1016	74,80	56,14	**78,39**
Design Aspect	100	45,24	55,43	52,38
	200	60,32	55,43	64,43
	500	63,25	55,43	67,95
	800	63,45	55,43	69,39
	1016	69,32	55,43	**73,52**
Customer Service Aspect	100	45,27	53,78	54,86
	200	54,58	53,78	74,58
	500	68,00	53,78	74,58
	800	74,16	53,78	77,46
	1016	79,07	53,78	**82,78**
Technology Aspect	100	40,00	55,85	47,50
	200	55,81	55,85	65,21
	500	63,64	55,85	70,61
	800	65,16	55,85	75,30
	1016	65,61	55,85	**77,90**

Table 5 shows that incorporating prior knowledge can help improve our previous best model (Logistic Regression). The model's performance on varying number of training set is evaluated on a fixed testing set for fairness/concistency reason. Initially, when the number of training data is small, the Logistic Regression gives result the lowest F1 score. Adding prior knowledge at this stage helps increase the inference process, although the performance is still below the rule based model. As the size of training data grows, the performance of both two models (model Logistic Regression and Logistic Regression with Prior Knowledge) increases. The Logistic Regression with Prior Knowledge however gives the better performance at any training size over the Logistic Regression model. The rule based model shows constant score as this is a deterministic function. Once the training data is sufficient, the Logistic Regression with Prior Knowledge achieves highest score. We then leverage the trained Logistic Regression with prior knowledge to be applied into full data so that the index can further be constructed based on the inferred results.

4.2 Result on the Index Series

After all the reviews are inferred (by Logistic Regression with prior knowledge), we then construct the index based on the polarity i.e. negative sentiment over all reviews on each particular NB_PSP's and specific time granularity. In this case, we discuss on two examples of the NB_PSP's (NB_PSP1 and NB_PSP2) having slightly different trends, yet still describing the increase of the digitalisation during the pandemic of COVID-19.

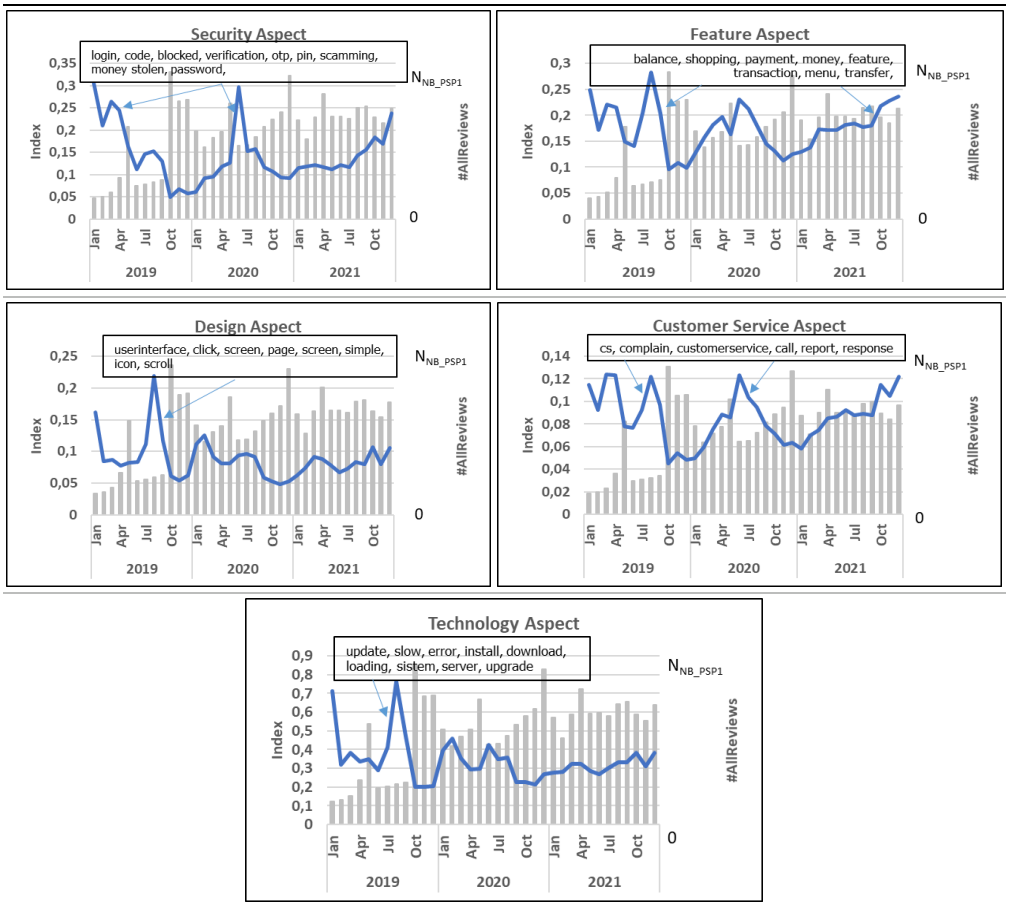
We investigate the series, mainly since 2019, because at this year the license for all the NB-PSP's was issued. In addition to that, the number of customers started to increase since 2019 as any financial activities were more becoming digitalised.

The series index for NB_PSP1 for 5 aspects can be seen in Fig 3 while NB_PSP2 for 5 aspects in Fig 4. The blue line (for NB_PSP1)/red line (for NB_PSP2) represents the index (left hand side), while the grey bar represents the total reviews (right hand side). For confidentiality reason, we denote the total reviews on the graphic as N_{NB_PSP1} (or N_{NB_PSP2}) for the number of review of NB_PSP1 (or NB_PSP2). The higher index indicating the more negative sentiments received. The period shown is in the monthly basis. The list of words on the annotation (on event analysis) represents the list of

topics being reviewed. They are based on the order of most frequent words extracted from the inferred negative sentiments.

Series of Index of NB_PSP1

Figure 3



* The period is from Jan 2019 until Dec 2021

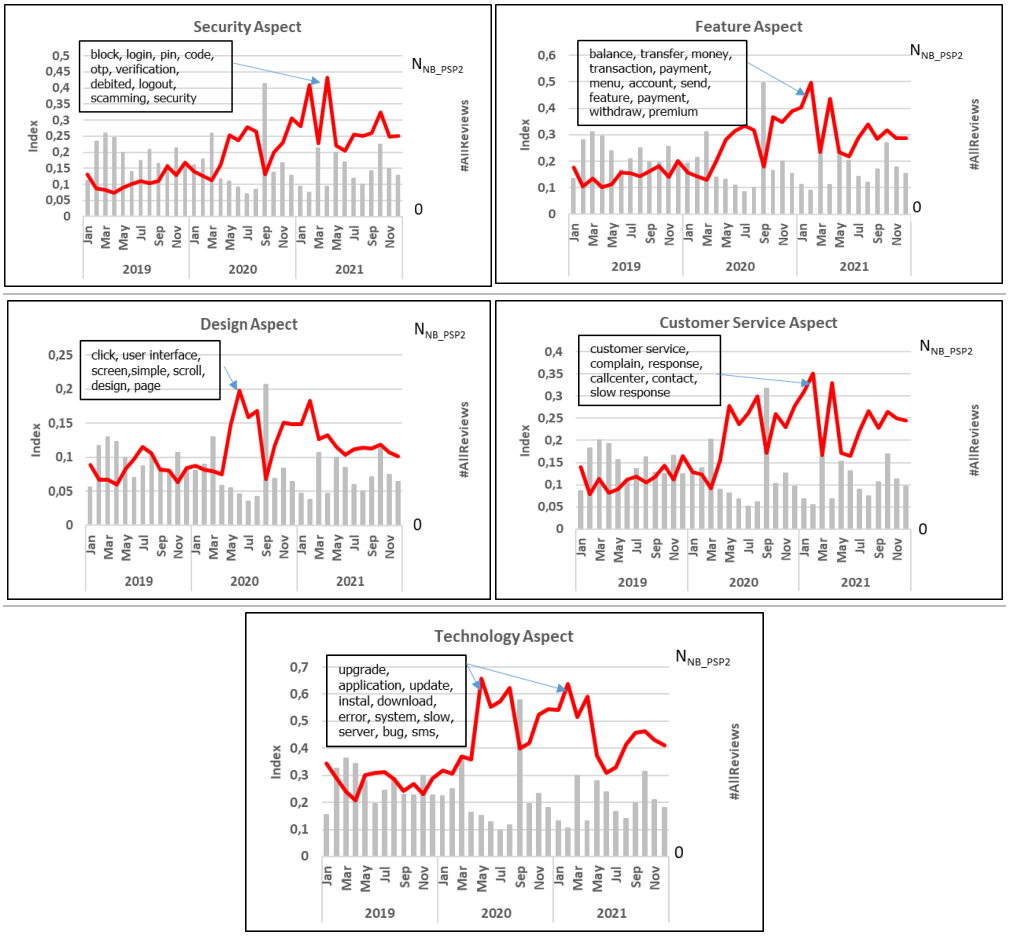
From Fig 3 we can see that during the period of 2019 until 2021, the number of reviews gradually increased. In March and April 2019, the NB_PSP1 showed the increasing index on Security aspect, Feature aspect, as well as Customer Service aspect. At this stage, a number of customers were identified to complain about problems on the e-money in the payment services provided. It was the period when the NB_PSP1 expansively promoted their service product after legally licensed by the authority.

In the period of July 2020 when the pandemic of COVID-19 had started to hit the country, the customer's complaint and negative sentiments were about the Security, Feature and Customer Service's aspect. At this period, the physical distancing and lock down policy by the government were conducted. The number of customers increased as most of them practiced the digital transactions, and thus the mobile payment apps were in very high demand. The Technology and Design aspects, however, showed only once dramatic increasing curve in July 2019. Afterwards, the number of customers complained about its technological and design aspect declined. It is identified that the NB_PSP1 made some improvements to enrich the customer experiences and apps reliability, especially during the pandemic of COVID-19.

On the other hand, the NB_PSP2 (Fig 4) depicted the relatively different trend. We can see that during the period of 2019 until 2021 the number of overall reviews showed relatively no significant increase, except in September 2020. The index, however, started to increase at the time the pandemic of COVID-19 hit the country. It is identified that the app's demand was highly increasing, and some possible trouble or unsatisfactory things were reported by user more than before. However, almost 5 aspects of NB_PSP2 showed relatively less differences of the trend, except in Q2 2020 as well as Q1 2021. In addition to that, in these two periods the Technology aspect was the most complained aspect.

Series of Index of NB_PSP2

Figure 4



* The period is from Jan 2019 until Dec 2021

5. Conclusion and Future Works

5.1 Conclusion

Through this work, we have performed the sentiment analysis method of user's reviews collected from the mobile application's store in order to oversee the quality of mobile payment apps of a nonbank-PSP. We conducted some experiments on the inference models: deterministic model (rule based), model-from-data (machine learning), and model-from-data incorporating prior knowledge. In our case study, the

model-from-data incorporated with prior knowledge helps infer the sentiment when the access to the training data is limited. Having demonstrating it onto 5 aspects, we suggest that this approach can be used to automatically infer the user's sentiment of mobile non-bank payment apps by some aspects.

We also propose an index based on the inferred sentiment and the number of overall reviews, that can represent the recent condition. By automatically analyse the sentiment and monitoring the series periodically we can suggest that this approach can be used in order to timely monitor the nonbank-PSP performance, e.g. as a leading indicator. Therefore, any risks on monetary and payment system affected by the issues in non-bank payment system environment hopefully can be anticipated in advance, as early as possible.

5.2. Future Works

We notice that there are still some improvements needed in our works. We thus highlight some future directions:

- **Advancement on Text Normalisation Methods**.
  In our model, we implement a deterministic approach by utilising regular expression. In order to improve the inference's performance, more sophisticated text normalisation leveraging more advanced language modelling should be used.
- **Enrich more various user's expression**.
  The method presented in this work relies on textual feature. In the future, the model should also be able to cover sarcastic text as well as emoji/emoticon.
- **Add more Human Languages**.
  Although still in a small portion, some users write review in non-Bahasa Indonesia, for example English. Thus in the future, the model should be able to tackle non-Bahasa Indonesia language, although its proportion is relatively small.

References

Arcand, M., PromTep, S., Brun, I. and Rajaobelina, L. (2017). "Mobile banking service quality and customer relationships", International Journal of Bank Marketing, Vol. 35 No. 7, pp. 1068-1089.

Bach, M.P., Starešini´c, B, Omazic, M.A., Aleksic, A. (2020). "m-Banking Quality and Bank Reputation". Sustainability 2020, Vol 12, pp. 4315.

Chawla, V.N., Bowyer, K. W., Lawrence, O.H.. 2002. Kegelmeyer, P.W. SMOTE: "Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research, Vol 16.

Fang, X., and Zang, J. (2015). "Sentiment analysis using product review data". Journal of Big Data vol. 2 No. 5.

Ganguli, S. and Roy, S.K. (2011). "Generic technology based service quality dimensions in banking: Impact on customer satisfaction and loyalty". International Journal of Bank Marketing, Vol. 29 No. 2, pp. 168-189.

Khan, A.G., Lima, R.P., Mahmud, M.S. (2021). "Understanding the Service Quality and Customer Satisfaction of Mobile Banking in Bangladesh: Using a Structural Equation Model". Global Business Review, Vol. 22 No. 1, pp. 85-100.

Leem, B.H. and Eum, S.W. (2021), "Using text mining to measure mobile banking service quality", Industrial Management & Data Systems, Vol. 121 No. 5, pp. 993-1007.

Rahman, A., Hasan, M., Mia, M. (2017). "Mobile Banking Service Quality and Customer Satisfaction in Bangladesh: An Analysis". The Cost and Management, Vol. 45 No. 2.

Rod, M., Ashil, J. Nicholas, Shao, J., Carruther, J (2009). "An examination of the relationship between service quality dimensions, overall internet banking service quality and customer satisfaction: A New Zealand study". Marketing Intelligence & Planning, Vol. 27 No. 1, pp. 103-126.

Sagib, G. K., & Zapan, B. (2014). "Bangladeshi mobile banking service quality and customer satisfaction and loyalty". Management and Marketing, Vol. 9 No. 3, pp. 331–346.

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).

Singla et al, 2017, "Sentiment Analysis of Customer Product Reviews using Machine Learning". Proceeding of 2017 International Conference on Intelligent Computing and Control (I2C2).

Thakur Rakhi, (2013). What keeps mobile banking customers loyal. "The International Journal of Bank Marketing", Vol. 32, Iss. 7, pp 628.

Vu, P.M., Nguyen, T.t., Hung, V.P., Nguyen, T.T. (2015). "Mining User Opinions in Mobile App Reviews: A Keyword-based Approach". Proceeding of 30th IEEE/ACM International Conference on Automated Software Engineering (ASE).

BANK INDONESIA

Sentiment Analysis of User's Reviews on Non-Bank Payment Service Apps[1]

N. Heru Praptono[2], Arinda D. Okfantia, M.Khoyrul Hidayat, M. Hafiruddin

Statistics Department

Basel, Switzerland
August 2022

[1] The expressed views belong to the authors and **do not** necessarily reflect the institution.

[2] Corresponding email: nursidik_hp@bi.go.id/pra.heru@gmail.com

BANK INDONESIA

Outline

BANK INDONESIA

Background

- More banking and payment activities are currently being conducted into mobile application platform than the traditional ones, legally provided by the authorised entities either bank or non-bank institutions.
- In Indonesia for example,
  - By the end of 2021, there have been at least 41 non-bank payment service providers (PSPs) who take a part in such banking and payment services.
  - The growing amount of non-bank PSP customers has found to be increasing recently, moreover since the pandemic of COVID-19.
  - Have infiltrated on nearly the whole of about 270 million people of the Indonesian population $\rightarrow$ play significant role in promoting and accelerating the economic growth.
  - Need to be supervised and monitored by the policy maker (i.e. Bank Indonesia) to anticipate any systemic risk.
- Measuring and monitoring the quality of non-bank payment service apps is difficult – but possible by considering user's review. Survey is costly, solution: apps review (i.e. in Google Play Store, Apps Store) as a proxy.
- Related works on mobile apps user's review analysis, e.g. Vu et al 2015 (mobile apps), Leem & Eum 2021 (m-banking).
- Some remaining issues (including but not limited to): non-formal text, limited training data, imbalanced training data, further utilisation for monitoring.

Overall Methodology

Scraping

Infer Sentiment by model

Index Calculation

Model Preparation

Data Annotation

Text Normalisation
utilizing regex and dictionary

Feature Extraction
tfidf bag of words

Learning & Inference

Evaluation

5 fold CV

Unnormalised Text (raw text)	Normalised Text (cleaned text)
Aplikasi yg Bagus.. Tampilannya menarik..	aplikasi yang bagus tampilannya menarik *(in English: nice application, yet interesting user interface)*

$$tfidf_{w,d} = tf_{w,d}\ x\ \log\frac{N}{df_w}$$

1. **Security**: Represents the security of mobile apps. some issue related to the problem on security aspects including log in/log out, registration, security of e-money.
2. **Feature**: Represents the variability and functionality of the feature offered by non-bank PSP's
3. **Design:** Represents how well the design related to the intuitiveness and user experiences.
4. **Cust. Services:** Represents how responsible, reliable, and effectiveness of the customer services.
5. **Technology**: Represents how advanced the technology the non-bank PSP's is, and represent how well the application is from the technological issues.

e.g.: Bach et al, 2020, Leem & Eum 2021

$$\text{idx}_{o,t} = \frac{\sum_i 1_{f(x_{0,t})=1}}{N_{o,t}} x \log N_{0,t}$$

Annotated Data and Model Evaluation

Annotated data are relatively imbalanced → utilise SMOTE (Chawla et al, 2002) during learning process.
Some experimented models → rule based model, SVM (with RBF Kernel), Decision Tree (Gini Splitting Criteria), and Logistic Regression. The experiment is performed with 5-fold cross validation.

Annotated Data

Annotated Dataset — Table 1

	Negative	Positive or Not Related	Total
Security Aspect	449	821	1270
Feature Aspect	420	850	1270
Design Aspect	306	964	1270
Customer Service Aspect	390	880	1270
Technology Aspect	427	843	1270

[1] Annotation is performed by 3 annotators.

Model Evaluation

Performance matrix of models from data (F1 in %) — Table 4

	Security Aspect	Feature Aspect	Design Aspect	Customer Service Aspect	Technology Aspect
Rule Based[1]	56,49	56,14	55,43	53,78	55,85
SVM	76,14	74,67	66,67	66,55	*65,77*
Decision Tree	72,00	61,90	63,37	59,36	63,69
Logistic Regression	*78,17*	*74,80*	*69,32*	*79,07*	65,61

[1]Rule based: $r(S) = \exists(\mathrm{kw}_{\mathrm{aspect}}) \text{ in } S \wedge \exists(\mathrm{kw}_{\mathrm{negative\_sentiment}}) \text{ in } S$

Incorporating Prior Knowledge

We adopt the methodology incorporating prior knowledge described by Schapire et al, 2002.

- **Prior Knowledge:** $r(S_x)$ (essentially the result of rule based model)
- **Learning**: 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" quantification, defined by:

$$\pi(x) = p(y = 1|x) = \begin{cases} 0.9; \text{ if } r(S_x) = 1 \text{ (true)} \\ 0.1; \text{ otherwise} \end{cases}$$

- **Inference**:

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

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

Incorporating Prior Knowledge (Cont'd)

Experimental Result

- The model's performance on varying number of training set is evaluated on a fixed testing set for fairness reason.
- Initially, when the number of training data is small, the Logistic Regression gives result the lowest F1 score.
- Adding prior knowledge at this stage helps increase the inference process, although the performance is still below the rule based model.
- As the size of training data grows, the performance of both two models (model Logistic Regression and Logistic Regression with Prior Knowledge) increases.
- The Logistic Regression with Prior Knowledge however gives the better performance at any training size over the Logistic Regression model.
- The rule based model shows constant score as this is a deterministic function.
- Once the training data is sufficient, the Logistic Regression with Prior Knowledge achieves highest score.
- We then leverage the trained Logistic Regression with prior knowledge to be applied into full data so that the index can further be constructed based on the inferred results.

Performance on Logistic Regression with Prior Knowledge (F1 in %)

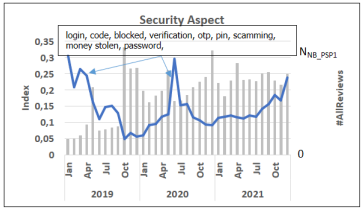
Table 5

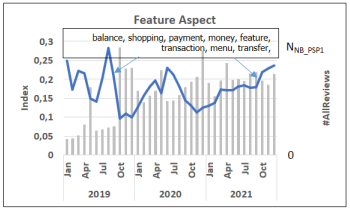
	# of Training Data	Logistic Regression	Prior Knowledge (Rule Based Model)	Logistic Regression with Prior Knowledge
Security Aspect	100	54,11	56,49	55,00
	200	66,67	56,49	67,86
	500	72,64	56,49	75,20
	800	75,76	56,49	78,82
	1016	78,17	56,49	**81,36**
Feature Aspect	100	50,39	56,14	57,65
	200	51,81	56,14	74,90
	500	62,50	56,14	77,67
	800	64,15	56,14	77,87
	1016	74,80	56,14	**78,39**
Design Aspect	100	45,24	55,43	52,38
	200	60,32	55,43	64,43
	500	63,25	55,43	67,95
	800	63,45	55,43	69,39
	1016	69,32	55,43	**73,52**
Customer Service Aspect	100	45,27	53,78	54,86
	200	54,58	53,78	74,58
	500	68,00	53,78	74,58
	800	74,16	53,78	77,46
	1016	79,07	53,78	**82,78**
Technology Aspect	100	40,00	55,85	47,50
	200	55,81	55,85	65,21
	500	63,64	55,85	70,61
	800	65,16	55,85	75,30
	1016	65,61	55,85	**77,90**

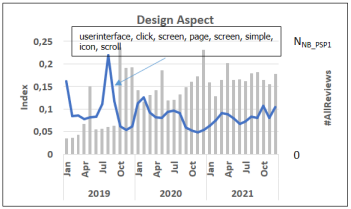
BANK INDONESIA

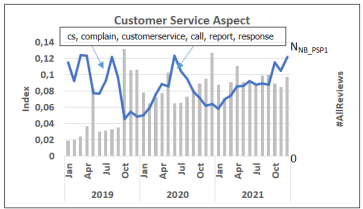
Result: Index (NB_PSP1)

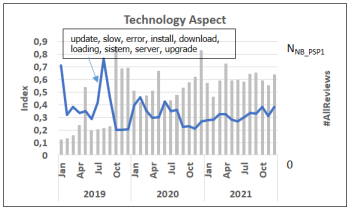
In March and April 2019, the NB_PSP1 showed the increasing index on Security aspect, Feature aspect, as well as Customer Service aspect. In the period of July 2020 when the pandemic of COVID-19 had started to hit the country, the complaint and negative sentiments were about the Security, Feature and Customer Service aspect.The Technology and Design aspects, however, showed only once dramatic increasing curve in July 2019 and declined afterwards → it was identified that some improvement happened.









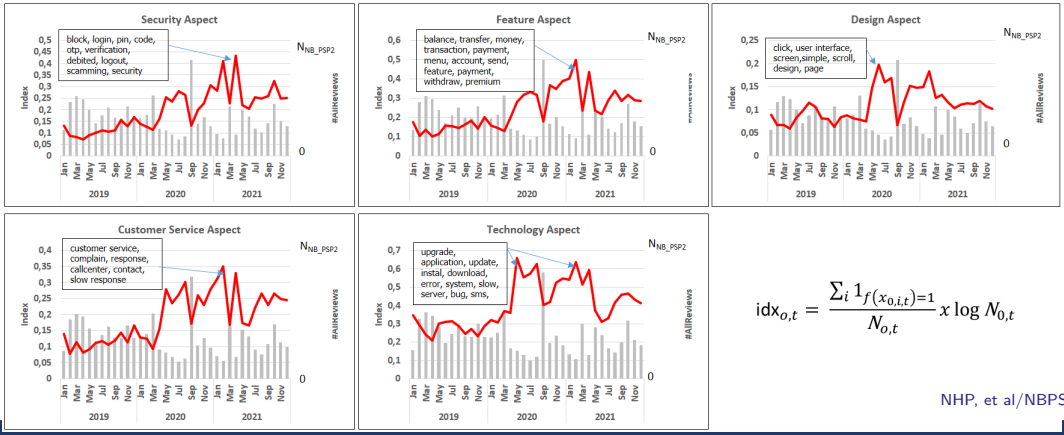


$$\text{idx}_{o,t} = \frac{\sum_i 1_{f(x_{0,i,t})=1}}{N_{o,t}} x \log N_{0,t}$$

BANK INDONESIA

Result: Index (NB_PSP2)

During the period of 2019 until 2021 the number of overall reviews showed relatively no significant increase, except in September 2020. The index, however, started to increase at the time the pandemic of COVID-19 hit the country. It is identified that the app's demand was highly increasing, and some possible troubles or unsatisfactory things were reported by user more than before. The most complained aspect was Technology aspect, 2020-Q2, 2021-Q1.



$$\text{idx}_{o,t} = \frac{\sum_i 1_{f(x_{0,i,t})=1}}{N_{o,t}} x \log N_{0,t}$$

BANK INDONESIA

Conclusion and Future Directions

Conclusion

1. We conducted some experiments on the inference models: deterministic model (rule based), model-from-data (machine learning), and model-from-data incorporating prior knowledge. Prior knowledge can help to improve the inference process when the number of training data is limited.
2. Having demonstrating it onto 5 aspects, we suggest that this approach can be used to infer the user's sentiment of mobile non-bank payment apps by some aspects in a big data quickly.
3. We also propose an index that is based on the inferred sentiment and the number of reviews. The series constructed by the index calculation represents the recent condition because the reviews can be scrapped at any time from the application's store.
4. By automatically analyse the sentiment and monitoring the series periodically we can suggest that this approach can be used in order to timely monitor the nonbank-PSP performance, e.g. as a leading indicator so that any further systemic risk can be anticipated in advance.

Future Directions

1. Advancement on Text Normalisation Methods.
2. Enrich more various user's expression.
3. Add more Human Languages.