Nhan Thanh Ngo

ngothanhnhan125@gmail.com

Abstract

This is technical note for the external review analysis including sentiment analysis and topic modelling

external review analysis

Technote

Revision

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Editor | Date | Change Note |
| 1 | nhan.ngothanh1 | 04/10/2022 | Initial |

Table of Contents

[Sentiment Analysis 3](#_Toc115967194)

[Executive Summary 3](#_Toc115967195)

[Goals 3](#_Toc115967196)

[Scope 3](#_Toc115967197)

[Implementation 3](#_Toc115967198)

[Folder Structure 5](#_Toc115967199)

[Result 6](#_Toc115967200)

[Code Storage 6](#_Toc115967201)

[Topic Modeling 7](#_Toc115967202)

[Executive Summary 7](#_Toc115967203)

[Goals 7](#_Toc115967204)

[Scope 7](#_Toc115967205)

[Literature Review 7](#_Toc115967206)

[Implementation 8](#_Toc115967207)

[Diagram 8](#_Toc115967208)

[Folder Structure 10](#_Toc115967209)

[Result 11](#_Toc115967210)

[Code Storage 14](#_Toc115967211)

[Reference 14](#_Toc115967212)

# Sentiment Analysis

## Executive Summary

### Goals

This model predicts sentiment for customers from their external and internal reviews to MyBANK 2.0 on both Android and iOS platforms. This supports management in getting the overview of customer interests and existing issues of MyBANK 2.0.

In this project there are two separate models for external reviews and internal reviews being built. The successful criterion is sentiment classification with accuracy at least 90% of total reviews based on model evaluation result.

### Scope

The model is built and applied based on external and internal reviews in the mobile apps. The model for external reviews could be trained with data enrichment from other bank review data.

## Implementation

The general framework for our analytics is generally described as follows.

**Text Processing (Vietnamese):** corrects the typo mistakes by looking up manually-defined dictionary.

**English Translation**: translates review after typo correction from Vietnamese to English via Google Translation API. Note: it takes long time for calling Google Translation API for large amount of data, ETA: 10 minutes/1000 requests.

**Text Processing (English):** includes removing emoji, words or letters which is not English, or in stop-word dictionary, or punctuations, followed by adjusting shortage writing like *can’t*, *didn’t*, etc…, and performing word lemmatization.

**Labeling**: Labeling (1: negative, 0: non-negative). It is the critical step, and actually hard to label exactly. I therefore propose labeling throughout voting scheme with three flags in order to increase the reliability of the labels. The flags are defined as follows.

* *Neg-pos Word Flag*: explores whether or not in review sentence there are negative and positive words. If only negative words appear, the flag would be True.
* *VADER flag*: VADER is the lexicon-based model which returns semantic score for sentence. Although the accuracy is not good enough for the Vietnamese-English translated reviews, however, its output is value to refer. The VADER compound score less than -0.05 would be considered as negative.
* *Rating flag*: rating score is a proxy for customer emotion to the app. Although such data might involve some biases, , the data are still meaningful and hence could be employed as a source for voting. The rule is once the rating less than or equal 3, the corresponding reviews would be considered as negative.

**Features**: the features include those from TF-IDF vectorizer (1000 features, ngram=(1,3)), neg-pos word flag, VADER flag. Those features are chosen because they extract important information of sentence directly. Those features are extracted from only text, no rating needed.

TF-IDF Vectorizer extracts sentence features used to train. Therefore, TF-IDF vectorizer runs before train test split. Because, test dataset also need output of TF-IDF to evaluate with the model.

**Model**: Training is implemented with 3 models, Naïve Bayes, Random Forest, and SVC. Accordingly, Naïve Bayes gives best performance in training and result response, however highest accuracy belongs to SVC. Therefore, SVC is chosen as final model. The model results are presented in following section.

A more illustrative process of the data processing can be seen in the following figures.

**BANK DB**

*Internal review*

**App (Android, iOS)**

*External review*

**Text Processing**

**(Vietnamese)**

**English Translation**

**Text Processing**

**(English)**

**Labeling**

Neg-Pos Word Flag

Rating Flag

VADER Flag

**Classification Model**

Naïve Bayes Multinomial

Random Forest

Support Vector Machine

**TF-IDF Vectorizer**

**Sentiment Model**

**Train Test Split**

**Train Model**

**Features**

TF-IDF Vectorizer

Neg-Pos Word Flag

VADER Flag

**BANK DB**

*Internal review*

**App (Android, iOS)**

*External review*

**Text Processing**

**(Vietnamese)**

**English Translation**

**Text Processing**

**(English)**

**Sentiment Model**

**Get result**

**Predicting**

**Load Model**

**TF-IDF Vectorizer**

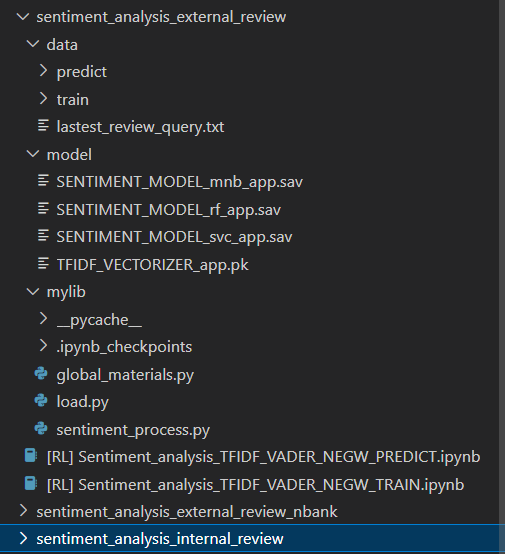
**Feature Extraction**

TF-IDF Vectorizer

Neg-Pos Word Flag

VADER Flag

## Folder Structure



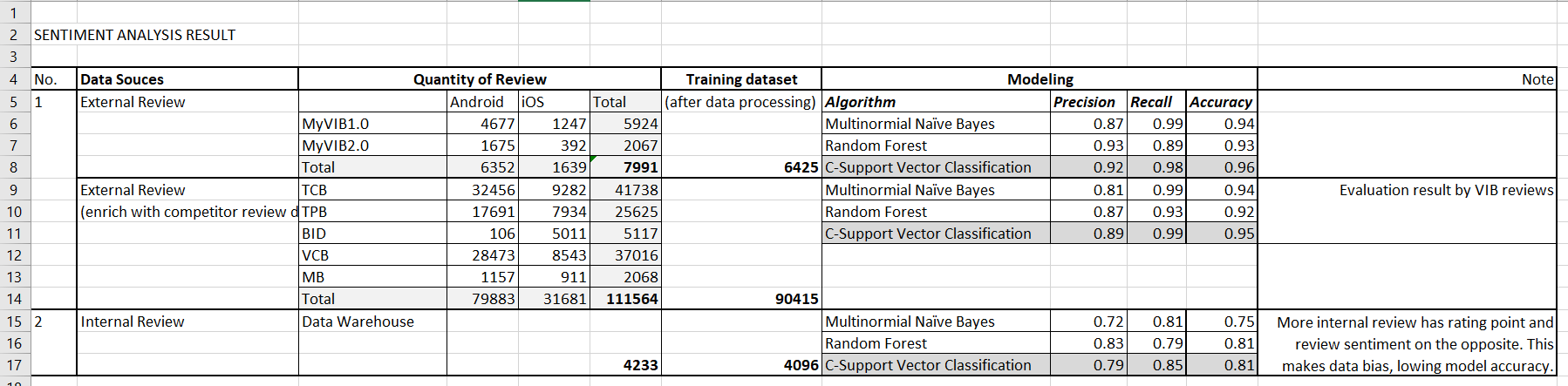
There are 3 sentiment models:

* Sentiment analysis for external review
* Sentiment analysis for external review enrich data of other banks
* Sentiment analysis for internal review

Each model has the folder structure as beside.

Model saved at “./model” folder

## Result



The SVC model returns the probability of review in [0,1], where 1 is negative, 0 is positive. The sentiment score of the review could be computed by **-2\*(probability-0.5).** The sentiment score is in range of [-1,1], where -1 stands for negative, and 1 for positive.

## Code Storage

In the code folder.

# Topic Modeling

## Executive Summary

### Goals

This model recognizes and distributes the reviews into relevant topics. This supports management in quick overview from a large number of reviews.

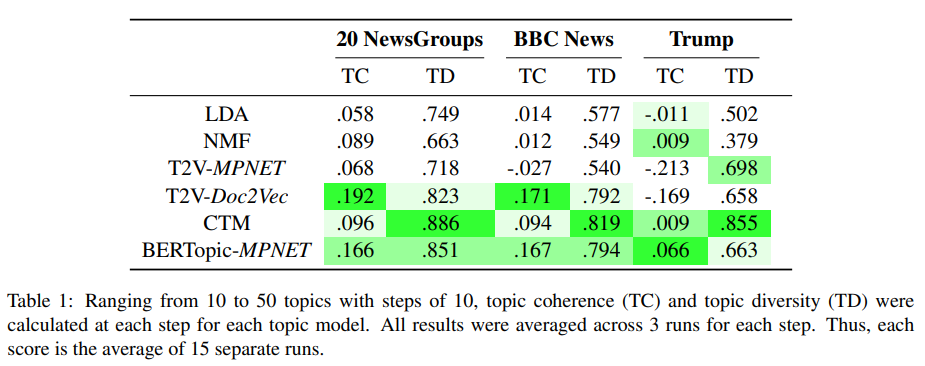
The conversion rate of current topic modelling is around 0.4 in last three months (7-9/2022). The target is improving at least 50% in this rate, accounts around 0.6.

### Scope

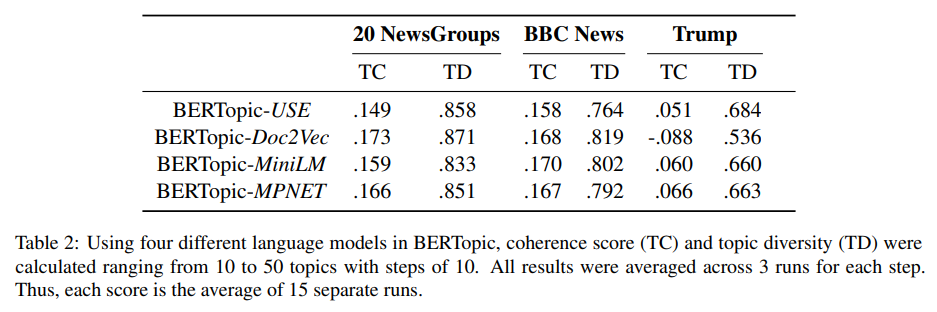
The model is built and applied based on external and internal reviews in MyBANK 2.0.

## Literature Review

As mentioned by Maarten Grootendorst [1], the comparison of the performance of various topic modelling models are shown in the Figure. Accordingly, BERTopic provides the better result in topic coherence (TC) and topic diversity (TD).



The research also compared the effectiveness of BERTopic with various sentence transformers. As shown in the figure, it is straightforward to get that BERTopic-miniLM and BERTopic-MPNET are a bit better than the others.



Based on the literature review, BERTopic with MiniLM or MPNET transformer is considered to use in this project.

## Implementation

### Diagram

The model includes two steps: training and predicting.

The training step involves feeding training dataset into the model for learning. After training and evaluation, the model and topic dictionary are saved, which would in turn be used in the prediction step.

The predicting part is tarted with loading saved topic model produced in previous step and running predictions for any reviews to get the topic categories.

Flow of Topic Model training and predicting could be refered at the following diagrams.

**Text Processing (Vietnamese):** corrects the typo mistakes by looking up the manually-defined dictionary. **English Translation**: translates reviews after typo correction from Vietnamese to English by Google Translation API.

**Text Processing (English):** includes removing emoji, words or letters which is not English, or from stop-word dictionary, or punctuations, followed by adjusting shortage writing like *can’t*, *didn’t*, etc.

**BERTopic Model** includes sequence of processing blocks:

* *Tokenizer*: using Count vertorizer with stopwords removing settings
* *Sentence Transformer*: using the weight “all-MiniLM-L6-v2” for fast training.
* *Dimension Reduction (UMAP):* using to reduce the number of sentence features when clustering.
* *Clustering (HDBSCAN):* clustering to figure out reviews which close each other.
* *c-TF-IDF*: using this algorithm, investigating to clusters in order to finding primary keywords representing in each cluster.

**Topic Regrouping:** the topic model return about 30 topics. There are some topics that could be grouped together further. Therefore, such resulting topics could be regrouped manually.

**BANK DB**

*Internal review*

**App (Android, iOS)**

*External review*

**Text Processing**

**(Vietnamese)**

**English Translation**

**Text Processing**

**(English)**

**Tokenizer**

**Sentence Transformer**

**Dimension**

**Reduction (UMAP)**

**Clustering (HDBSCAN)**

**c-TF-IDF**

**keyword extraction**

**BERTopic Model**

**TopicModel Saved**

**Topic Regrouping**

**Negative Sentiment Reviews (train dataset)**

**Train Model**

**BANK DB**

*Internal review*

**App (Android, iOS)**

*External review*

**Text Processing**

**(Vietnamese)**

**English Translation**

**Text Processing**

**(English)**

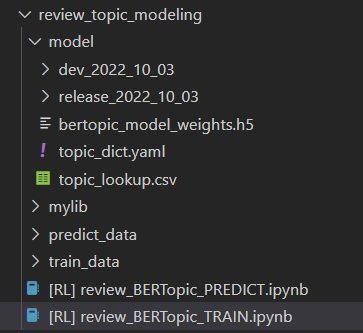
**TopicModel**

**Get result**

**Predicting**

**Load TopicModel**

## Folder Structure



* **Training:** *[RL] review\_BERTopic\_TRAIN.ipynb*
* This Training inherits the result external and internal sentiment.
* [Note] Run sentiment analysis for external and internal reviews before this BERTopic trains.
* Just click Run All, the result of model is saved at dev\_<date> folder. After checking the new version model in dev\_<date> folder, if it matches requirement, the folder could be changed to release\_<date> manually.
* **Predicting:** *[RL] review\_BERTopic\_PREDICT.ipynb*
  + Paramater: Adjust the parameter before running

# PARAMETER

TOPIC\_MODEL\_PATH = './model/release\_2022\_10\_03/'

DATE\_FROM = datetime.datetime(2022,9,1)

DATE\_TO = datetime.datetime(2022,9,30)

REVIEW\_FILE\_PATH = "../sentiment\_analysis\_external\_review\_nbank/data/predict/OUT\_sentiment\_prediction.csv"

TOPIC\_MODEL\_PATH: which model version you want to use

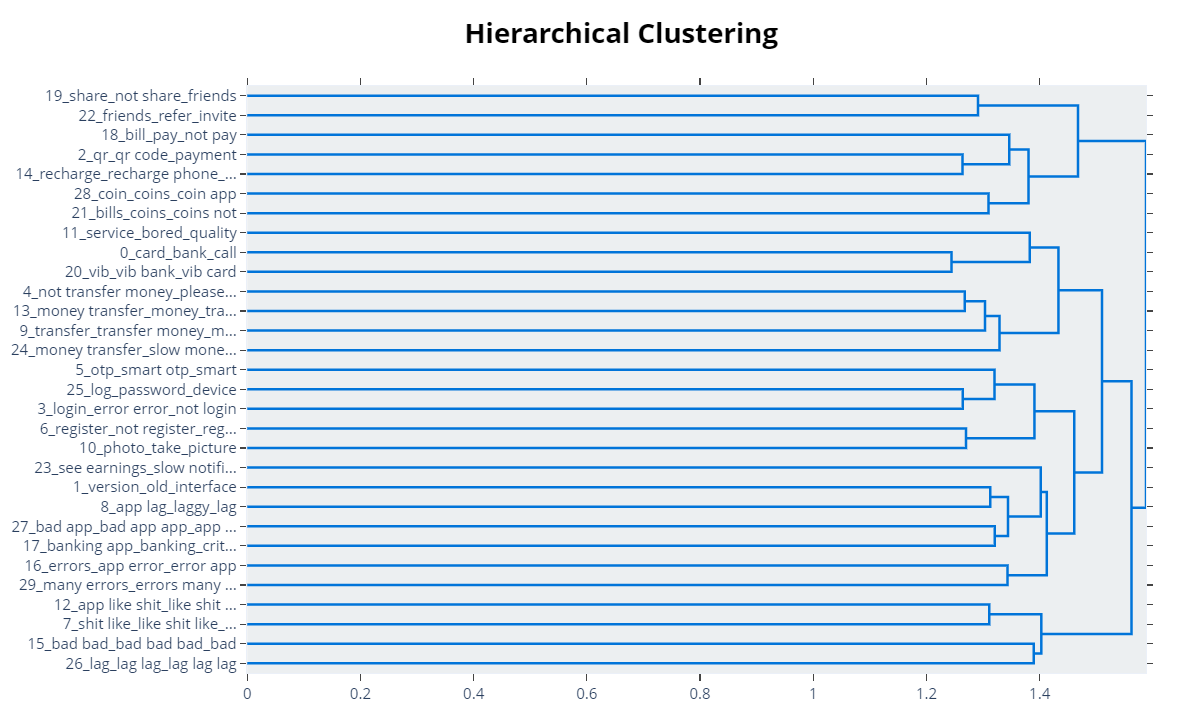
DATE\_FROM, DATE\_TO: reviews in date range will be used

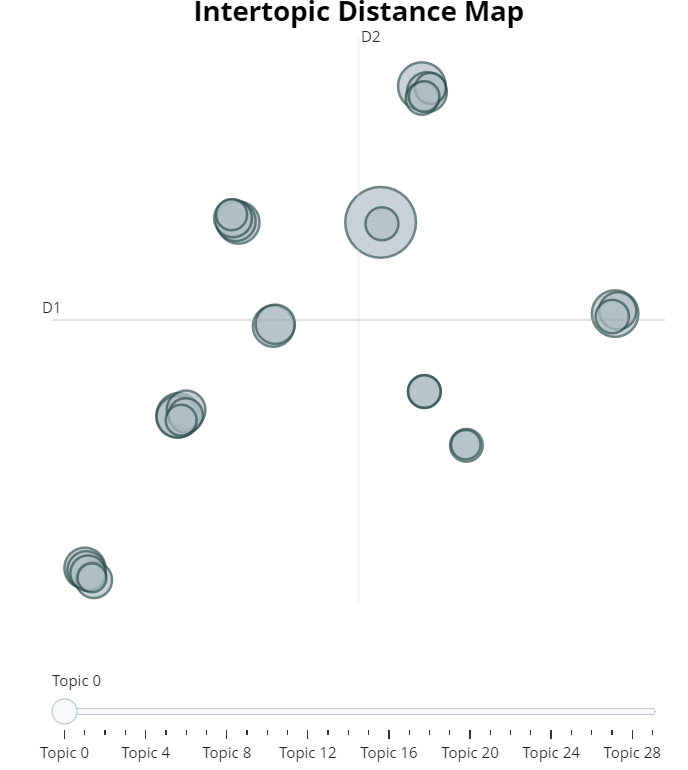
REVIEW\_FILE\_PATH: path to reviews file that you need to run topic modeling.

* **OUTPUT** **at** ./predict\_data/ OUTPUT\_topic\_modeling.csv

## Result

The conversion rate of the topic model is 0.6.



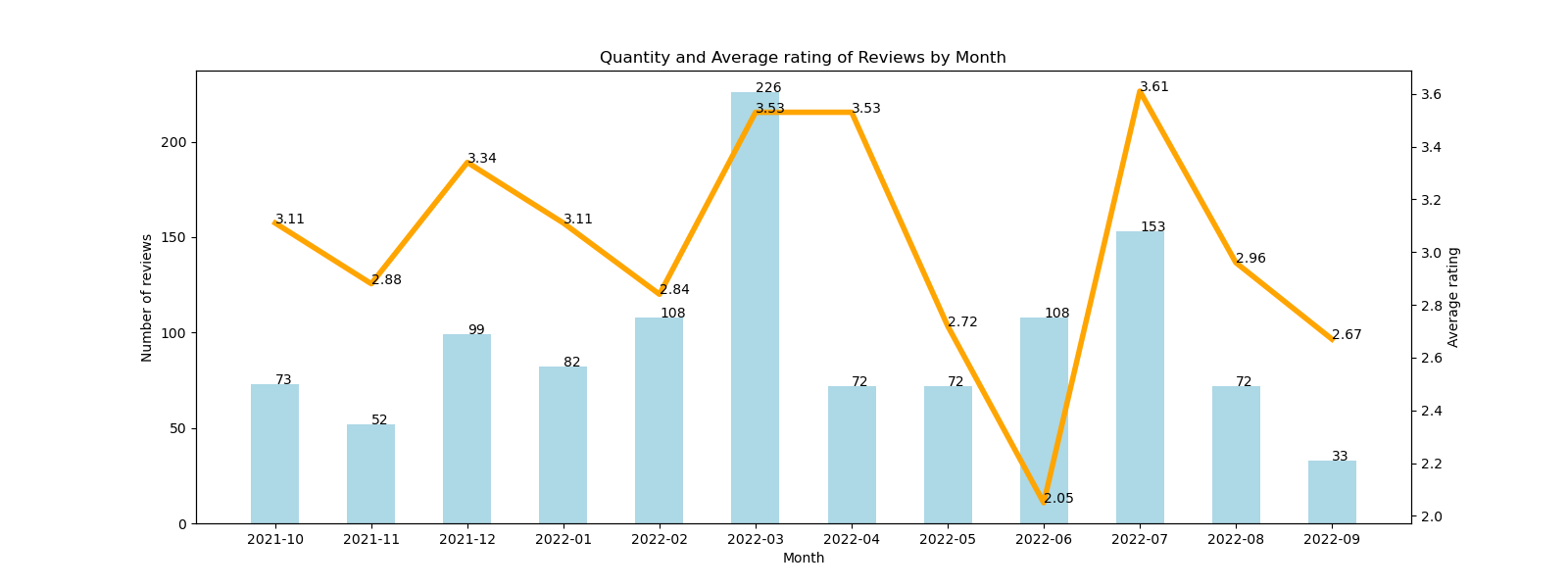


**Manual group topic result:** From 30 topics to 11 topics

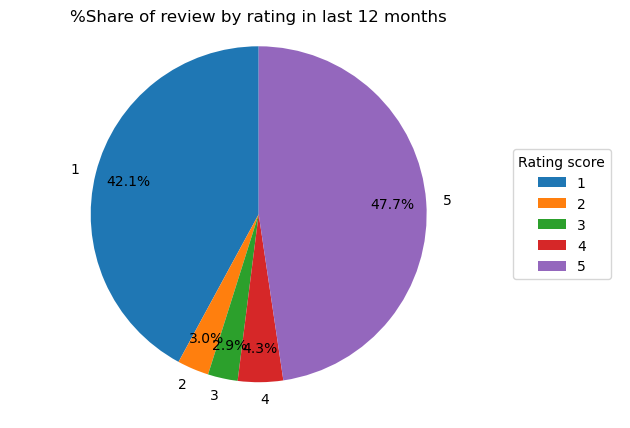
|  |  |  |
| --- | --- | --- |
| No. | Topic | Description |
| 1 | Issues relating to Coin Reward Program | Issues relating coins reward by introducing friends get coins, marketing campaigns, etc. such as slow receiving, losing coins. |
| 2 | Issues relating to Use/Payment by coins | Issues rating using coins to make payment. |
| 3 | Topup/Bill payment issues | Payment error, or can not implement when using topup/bill (eg. electricity, water) payment. |
| 4 | Issues relating to Money Transfer | Can not transfer/slow transfer/transfered but not received. |
| 5 | Login App/smart OTP issues | Error when login app, can not login  Wrong password when login or OTP confirm. |
| 6 | Hard eKYC registration issues | Issues relating to eKYC account registration, eg. face scanning not pass, ID card image fail. |
| 7 | Quality of app | Bad experience in app (slow/lag/unattractive) |
| 8 | App Error complaint | Errors relating to the app in general |
| 9 | Expressing disparage/disappoint to the app | Furious reviews about app. |
| 10 | Supporting staff complaint | Ineffective supporting staff such as slow response, irrelevant response, attitude. |
| 11 | Card issues | Issues relating to cards |

Below is the result of Digital Monthly Report for MyBANK10, external data in Android.

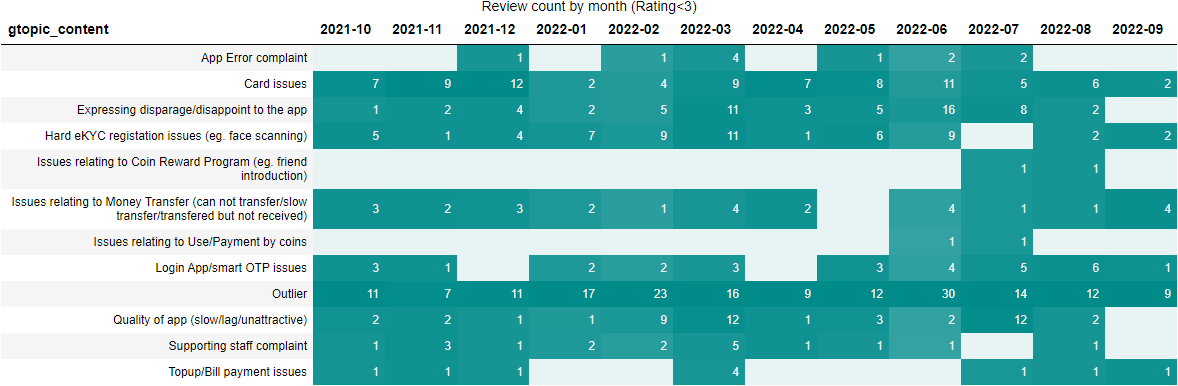
* **Figure 1:** Quantity and Average rating of reviews by Month



* **Figure 2:** %Share of review by rating in last 12 months



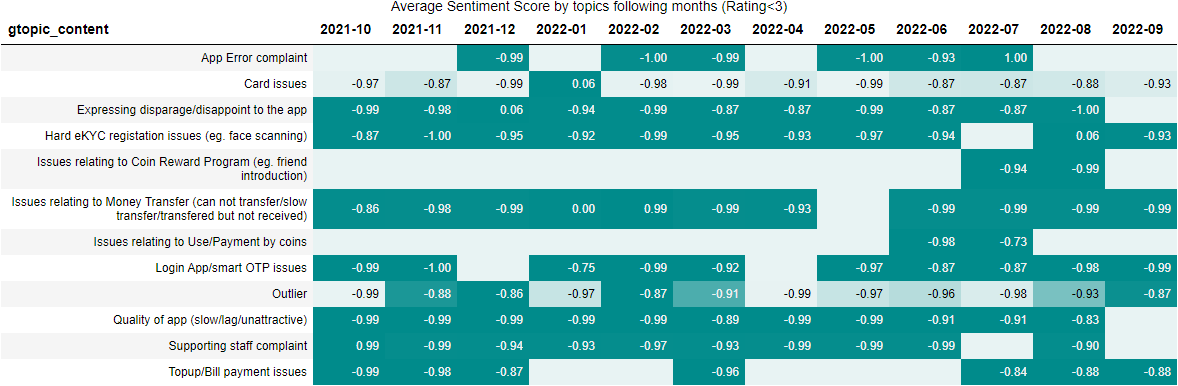
* **Figure 3:** Review count by months (Rating <3)



* **Figure 4:** Average Review rating by months (Rating <3)



* **Figure 5**: Average Sentiment Score by topics following months (Rating <3)



## Code Storage

In the code folder

## Reference

[1] Maarten Grootendorst. 2022. *BERTopic:* *Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794v1 [cs.CL]