BT5153
Applied Machine Learning for Business Analytics
(Final Project)

Singapore's Mobile Banking Applications:
An Analysis of User Reviews

Kenneth Leung Tin Yau (A0053782B)

Koh Lin Hui (A0206674B)

Nyoe Hwee Kian Sandy (A0206454J)

Pamela Chan Ting Jun (A0206466A)

Wong Choon Chuen (A0075374Y)

Singapore's Mobile Banking Applications: An Analysis of User Reviews

Kenneth Leung Tin Yau (A0053782B), Koh Lin Hui (A0206674B), Nyoe Hwee Kian Sandy (A0206454J), Pamela Chan Ting Jun (A0206466A), Wong Choon Chuen (A0075374Y)

Abstract

This paper examines the user reviews of Singapore's mobile banking applications ("apps"), to gain insights on user sentiments and what specific areas do users like and dislike about the mobile banking apps.

1. Introduction

Mobile banking in Singapore has been on the rise over the past three years. A 2017 HSBC report ("Why only 9% of Singaporeans", 2017) found that only 9% of Singaporeans preferred to use banking applications ('apps') on their mobile devices, due to concerns with cyber security. This figure rose sharply in 2018, despite many users reporting usage issues such as login difficulties and long loading times ("Singaporeans ditch bank visits", 2018). By 2019, 65% of bank customers in Singapore were using mobile banking apps (Ng, 2019). Customers who shun mobile banking included the elderly and individuals concerned with the app's security ("Cash is still king", 2019).

With the Singapore government's push to transform the country's financial technology (FinTech) ecosystem (Au Yong, 2016), citizens need to embrace digital banking technology. Increasing the adoption rates of mobile banking would be a good starting point. An incentive to encourage more bank customers to switch to mobile banking, is to improve app user interface and user experience. This is especially pertinent in Singapore as Singaporean customers are known to expect particularly high customer service standards (Tan, 2017).

Banks should also be incentivised to improve their apps given the myriad of consumer choices available in Singapore, where on average a Singaporean has about five different bank accounts (Ng, 2019). Competition in the banking services sector would also intensify going forward, with the issuance of five digital bank licenses by the Monetary Authority of Singapore (MAS) by mid-2020 (Choo, 2020). Strong FinTech contenders such as Grab-Singtel and Ant Financial were amongst those vying for these licenses. Competing banks will have to deliver on the user's digital banking experience to maintain their edge.

2. Scope

This study thus aims to gain insights on user sentiments towards mobile banking apps in general, and what specific areas do users like and dislike about these apps. This will be done through a deep-dive analysis into the online user reviews of various mobile banking products offered by banks in Singapore, from the Google Play Store and Apple App Store. The scope of study will cover 8 banks that specifically catered to Singapore customers.

We will employ Natural Language Processing (NLP) applications to extract useful information from the reviews, including unsupervised sentiment analysis techniques such as Stanford CoreNLP, and topic modeling based on Latent Dirichlet Allocation (LDA). We will later also use supervised classification models such as Multinomial Naïve Bayes to extract more useful insights from the data.

3. Dataset

3.1 Data Source

The online reviews were drawn from the two main sources of mobile app downloads in Singapore, the Google Play Store and the Apple App Store. We selected 8 banks' mobile banking apps with at least 100 reviews, for analysis.

- **DBS Bank:** DBS digibank SG and POSB digibank. POSB is grouped with DBS as it was acquired by DBS and operates as part of DBS.
- **Citibank:** Citibank SG
- **United Overseas Bank (UOB):** UOB Mighty Singapore
- **HSBC:** HSBC Singapore
- **Overseas-Chinese Banking Corporation, Limited (OCBC Bank):** OCBC SG Mobile Banking
- **Standard Chartered (SC):** SC Mobile Singapore
- **Maybank:** Maybank2u (Old) and Maybank2u (New). The latter was a new version of the app by Maybank and the older version was not phased out yet as of data extraction during February 2020.
- **CIMB:** CIMB Clicks Singapore

3.2 Data Scraping

User reviews were scraped from both the Google Play Store and Apple App Store.

- While the websites' source codes were in HTML, the data was retrieved by Java.
- There was a need to automate scrolling to the end of the page before more reviews were loaded. On the Google Play Store, a 'Show More' button needed to be clicked every 200 reviews.
- For longer reviews, a 'Full Review' button needed to be clicked before the full review would be loaded. However, the partial review and full review were stored in different HTML tags, therefore there was no need to automate the clicking on 'Full Review' button; the full review was simply scraped from its relevant HTML tag.

Selenium and BeautifulSoup4 were used for scraping all 10 mobile banking apps. Selenium was first used to automate an internet browser, for scrolling to the end of the page (for both Google Play Store and Apple App Store) and for clicking on the 'Show More' button (on Google Play Store). After all reviews were loaded, BeautifulSoup4 was then used to scrape the data under the relevant HTML tags.

3.3 Data Variables

A total of 5 variables were scraped:

- **Bank:** The name of the bank who owns the app.
- **Name:** The name of the user who provided the review. However, these are not unique ids.
- **Date:** The date that the user provided the review.
- **Number of Stars**: The number of stars that the user rated the mobile app (which can range from 1 to 5). On the Google Play Store, the number of stars were tagged as: 1 – Hated it; 2 – Didn't like it; 3 – Just OK; 4 – Like it; 5 – Loved it. There were no tags to the stars on the Apple App Store, although it can be assumed that it signifies similar meanings as the Google Play Store's stars.
- **Review:** The text content of the review.

3.4 Data Pre-Processing

A total of 24,760 user reviews from the Google Play Store and 7,695 user reviews from the Apple App Store were scraped. Some basic pre-processing steps were performed:

- The reviews from both stores were combined as they have the same 5 variables.
- The reviews from different apps but the same bank were combined (i.e. DBS digibank SG and POSB digibank, Maybank2u and Maybank2u (New)).
- Dropped 15 user reviews without review text

The above resulted in a dataset consisting of 32,440 user reviews and the 5 variables as outlined in section 3.3.

4. Exploratory Data Analysis

4.1 Number of Reviews

The three banks with the most reviews are DBS, OCBC, and UOB, consisting of 75.7% of the dataset (Figure 1). These are also the three largest banks by assets in Singapore (Lam, 2019).

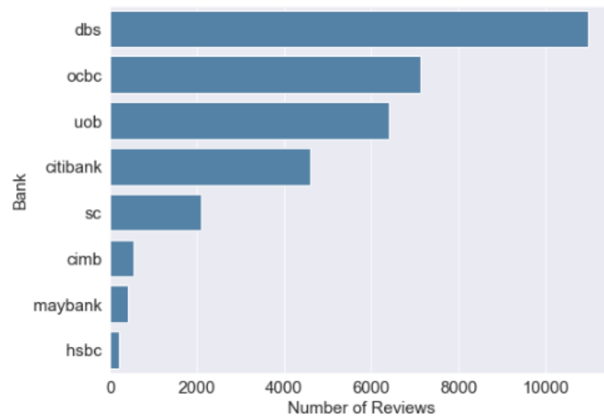


Figure 1. Number of app reviews for each bank.

4.2 Proportion of Stars

Overall, most of the reviews were 1- (59.6%) and 5-stars (17.1%) (Figure 2). Polarised reviews could also be found in most online reviews such as on Yelp and Amazon (Klein et al., 2018). All banks had more than 50% 1-star reviews. This is worrying as studies have shown that user satisfaction and purchase intentions are negatively influenced by negative reviews (Weisstein et al., 2017). Citibank had the most positively rated app, with the highest proportion of 4- and 5-star reviews (32.8%). This ties in with it being one of the best banks for customer satisfaction across the retail banking industry (Badenhausen, 2019).

4.3 Top Trigrams

The top 20 trigrams in reviews with stopwords (Figure 3) revealed that reviews were mostly negative, with phrases such as "please fix it" and "does not work". A positive "easy to use" was the second most common trigram, despite most reviews being negative. Stopwords would not be removed for sentiment analysis during feature engineering later, given that removing stopwords such as "not" could change the meaning of a sentence.

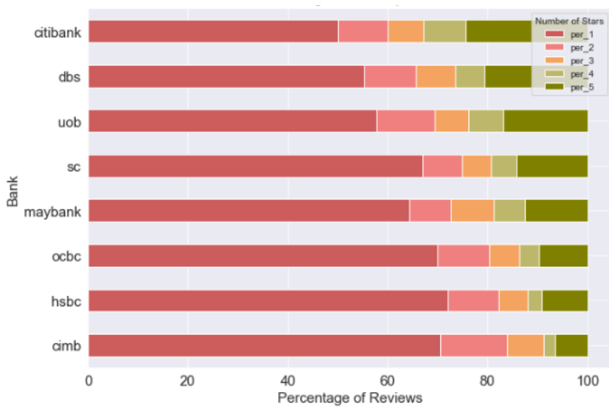


Figure 2: Proportion of Reviews by Stars.

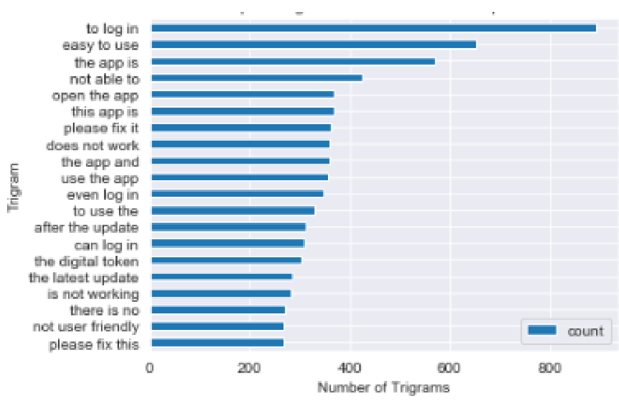


Figure 3: Top 20 Trigrams in Review (with stopwords).

5. Feature Engineering

New features were created for the study.

5.1 For Unsupervised Sentiment Analysis

5.1.1 STAR SENTIMENT

For the purpose of unsupervised sentiment analysis, a new variable *Star_Sentiment* is created. User reviews are categorised based on the number of stars ('Negative': 1- and 2-stars; 'Neutral': 3-stars; 'Positive': 4- and 5-stars).

The proportion of negative reviews (1- and 2- stars), neutral reviews (3-stars), and positive reviews (4- and 5-stars) were 69.9%, 7.0%, and 23.1% respectively.

5.1.2 PROCESSED REVIEW

A new variable *Processed_Review* was created through pre-processing the reviews:

- Removed punctuation and non-ASCII characters
- Lemmatised

5.2 For Topic Modeling

5.2.1 TOKENISED REVIEW

A new variable *Tokenised_Review* was created through pre-processing the reviews:

- Removed new line characters, single quotes, and non-alpha characters, punctuation, stopwords
- Converted to lower-case
- Tokenized
- Lemmatised

5.3 For Supervised Classification Models

5.3.1 VECTORISED REVIEW

Vectorised_Review was obtained by vectorising the *Processed_Review* using the Term Frequency-Inverse Document Frequency (TF-IDF), which accounts for the number of times each word appears in the review, and the inverse frequency of the word across all reviews.

5.3.2 PADDED REVIEW

For the Convolutional Neural Network (CNN) model, the *Processed_Review* was further processed to retain only the most frequent 7,999 words. It was then padded to the maximum length of all documents. This resulted in the new variable *Padded_Review.*

6. Unsupervised Sentiment Analysis

In this section, we explore user sentiments in reviews, and if the sentiments in user reviews were comparable to the sentiments roughly signified by the number of stars rated by users (*Star_Sentiment)*. Three unsupervised sentiment analysis algorithms were used to classify the user reviews to Negative, Neutral, and Positive sentiments (defined as *Review_Sentiment)*. The *Review_Sentiment* were then tested against the *Star_Sentiment* variable.

6.1 Models

6.1.1 TEXTBLOB

TextBlob's sentiment property returns a polarity score between -1 (negative) to 1 (positive). There are two sentiment analysis implementations available, one using PatternAnalyzer (using the Pattern library by De Smedt and Daelemans (2012)), and the NaiveBayesAnalyzer (using a NLTK classifier pre-trained on a movie reviews corpus by Pang & Lee (2005). Both analyzers were tested using the *Processed_Review* dataset as it is a lexicon-based model.

The NaiveBayes Analyzer only outputs Positive or Negative sentiments, thus we observe that the precision, recall, and F1-score of the Neutral class using the NaiveBayesAnalyzer were 0.00. Nonetheless, by comparing the three performance metrics for the Positive and Negative classes of these two analyzers, the PatternAnalyzer outperforms the NaiveBayesAnalyzer in our case. As such, the PatternAnalyzer was selected as the implementation of choice for TextBlob in our study.

Table 1: Classification report for TextBlob (Pattern).

	PRECISION	RECALL	F1-SCORE	SUPPORT
NEGATIVE	0.91	0.39	0.55	22,644
NEUTRAL	0.08	0.41	0.13	2,264
POSITIVE	0.51	0.77	0.61	7,532
OVERALL	0.50	0.52	0.43	32,440

Table 2: Classification report for TextBlob (NB).

	PRECISION	RECALL	F1-SCORE	SUPPORT
NEGATIVE	0.81	0.44	0.57	22,644
NEUTRAL	0.00	0.00	0.00	2,264
POSITIVE	0.30	0.80	0.44	7.532
OVERALL	0.37	0.42	0.34	32,440

6.1.2 NLTK VADER

The Natural Language Toolkit (NLTK)'s Valence Aware Dictionary and sEntiment Reasoner (VADER) developed by Hutto & Gilbert (2014) was used to assign every user review a compound sentiment score. The score, ranging from -1 (most negative) to 1 (most positive), is calculated by summing the valence scores of each word, adjusted according to the rules, and then normalised.

The reviews are then classified based on their compound sentiment score ('Negative': score ≤ -0.05; 'Neutral': -0.05 < score < 0.05; 'Positive': ≥ 0.05). The algorithm's advantage is that it was developed for social media texts.

The NLTK VADER was best able to predict negative reviews, with a precision of 0.92. See Table 3 for the classification report. It. However, the precision, recall, and F1-score of the Neutral class remains low. As the algorithm is based on lexicon meanings, it could mean that users who provide neutral ratings were giving review texts which contains words with a certain positive or negative polarity.

Table 3: Classification report for NLTK Vader.

	PRECISION	RECALL	F1-SCORE	SUPPORT
NEGATIVE	0.92	0.44	0.60	22,644
NEUTRAL	0.08	0.29	0.12	2,264
POSITIVE	0.46	0.80	0.58	7,532
OVERALL	0.48	0.51	0.43	32,440

6.1.3 STANFORD CORENLP

We next explored the Stanford CoreNLP algorithm, which is implemented based on Socher et al's (2013) sentiment model. Its advantage is that it can account for sentiments of word phrases, beyond just lexicons. Like TextBlob's NaiveBayesAnalyzer, it was trained on the same movie reviews corpus by Pang & Lee (2005). The algorithm classifies reviews based on a sentiment value ranging from 0 (very negative) to 4 (very positive).

The model was first applied on the unprocessed review dataset. The output was grouped as negative (0 and 1), neutral (2), and positive (3 and 4), before being tested against the *Star_Sentiment.* The classification of Neutral reviews remained weak, with a precision of 0.07.

Table 4: Classification report for Stanford CoreNLP (on unprocessed reviews).

	PRECISION	RECALL	F1-SCORE	SUPPORT
NEGATIVE	0.85	0.63	0.72	22,644
NEUTRAL	0.07	0.31	0.11	2,264
POSITIVE	0.69	0.52	0.59	7,532
OVERALL	0.54	0.49	0.48	32,440

The model is further tested on the *Processed_Review* dataset. See Table 5 for the classification report. Overall, precision, recall, and F1-scores of Stanford CoreNLP performed better than TextBlob and NLTK VADER.

Table 5: Classification report for Stanford CoreNLP (on processed reviews).

	PRECISION	RECALL	F1-SCORE	SUPPORT
NEGATIVE	0.87	0.65	0.74	22,625
NEUTRAL	0.08	0.35	0.14	2,261
POSITIVE	0.69	0.56	0.62	7,464
OVERALL	0.55	0.52	0.50	32,350

6.2 Model Comparison

A final comparison of the models' test F1-scores is shown in Table 6. The F1-score was chosen as the metric for comparison instead of accuracy, as there is a clear presence of skewed classes in our dataset, where there are many more reviews with negative sentiment as compared to the other sentiments.

From the comparison, we can see that the sentiments generated by Stanford CoreNLP on the *Processed_Review* gave the best performance. As such, the sentiments from this model was used for our *Review_Sentiment* feature.

Table 6: Test F1-scores of unsupervised sentiment analysis models.

Model	Positive	Neutral	Negative
TextBlob (Pattern)	0.61	0.13	0.55
NLTK VADER	0.59	0.12	0.60
Stanford CoreNLP (Unprocessed)	0.59	0.11	0.72
Stanford CoreNLP (Processed)	0.62	0.14	0.74

6.3 Results Analysis

6.3.1 Comparison of Sentiments

The results showed that a large proportion (about 40%) of the dataset had differing *Review_Sentiment* and *Star_Sentiment*. This discrepancy between sentiment scores and star ratings has also been detected in other studies. Lak & Turetken (2014) attributed it to the tendency of people to use more neutral language when expressing their opinions in natural language. Our findings (Figure 4) support this hypothesis, as 79.9% and 66.7% of the wrongly classified negative and positive *Star_Sentiment* were classified as neutral *Review_Sentiment*.

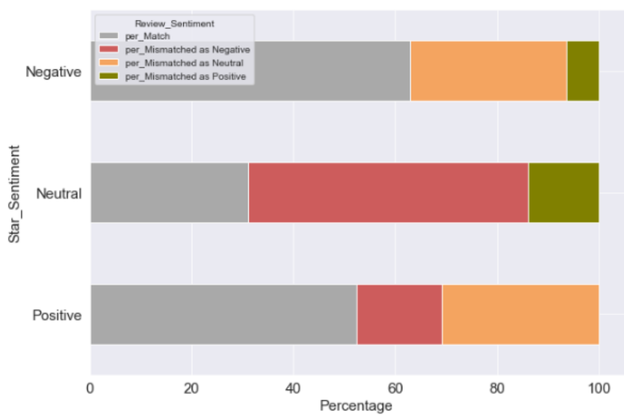


Figure 4: Comparison of *Star_Sentiment* and *Review_Sentiment.*

It was also observed that 78.0% of the wrongly classified neutral *Star_Sentiment* were classified as negative *Review_Sentiment*. This suggests that users might be leaving reviews with harsher sentiments than intended, or focusing more on the negative aspects of the app.

About 20% of the wrongly classified positive and negative *Star_Sentiment* were classified as being on the opposite spectrum with CoreNLP (i.e. positive *Star_Sentiment* classified as negative *Review_Sentiment*, and vice versa). In the former, users may be again writing only on the negative aspects of the app in their review; while in the latter, users may have been again been using more neutral language/ more lenient when writing the review.

6.3.2 For Deeper Analysis

It was found that there were discrepancies between the *Review_Sentiment* and *Star_Sentiment*. The discrepancy could be because every user has different standards when translating their sentiments to a pre-defined set of stars. Therefore, the banks should not solely rely on the number of stars to measure user satisfaction, but also perform a deeper analysis of reviews, such as topic modeling to find out the areas of satisfaction or dissatisfaction. This is explored in Section 7 of this paper.

Another resultant area of interest was if *Star_Sentiment* could be predicted through other text patterns (non-sentiment analysis) for the 40% of the dataset where *Review_Sentiment* does not match *Star_Sentiment.* If it could not be classified through both unsupervised and supervised models, it might imply that there were other external non-textual attributes that affected user ratings, such as cases of spam reviews, users clicking wrongly, etc. This is explored in Section 8 of this paper.

6.4 Business Insights

Through conducting unsupervised sentiment analysis, banks could have a rough sense of the sentiments of their users. For instance, a word cloud of the negative comments overall (Figure 5) showed that most reviews were about login and security token issues, while a word cloud of the positive comments overall (Figure 6) showed that users were happy about cases such as their bank app providing support for syncing with smartwatches.

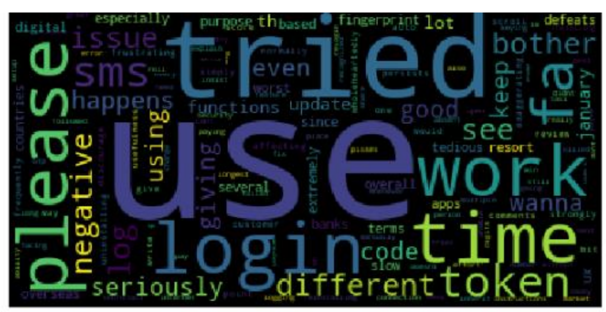


Figure 5: Word Cloud of all Negative Reviews.

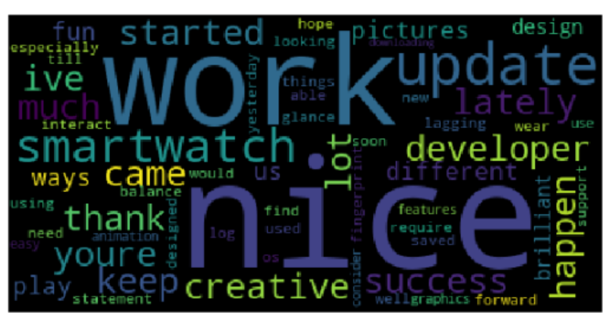


Figure 6: Word Cloud of all Positive Reviews.

6.4.1 DBS

We took a closer examination into the reviews of DBS, for examples of actionable business insights that could be derived through topic modeling. DBS was chosen here as they had the highest number of reviews amongst all the banks. The word clouds for DBS (Figures 7 and 8) showed that users for DBS were generally having security issues such as with their digital tokens and 2 Factor-Authentication (2FA) login. They were also happy about their DBS app syncing with smartwatch. The bank could put in more app development to improve the security aspect of the app. See Appendix A for the word clouds for the other 2 large banks, OCBC and UOB.

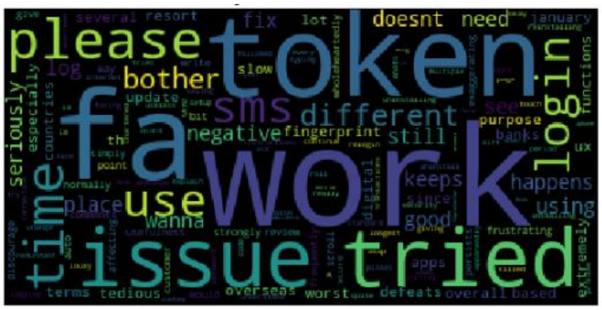


Figure 7: Word Cloud of all Negative Reviews for DBS.

Figure 8: Word Cloud of all Positive Reviews for DBS.

7. Topic Modeling

As established in section 6.3.3, there was a need for a deeper analysis of reviews to find out areas of satisfaction or dissatisfaction of users. This section details the steps to segment user reviews to specific positive and negative topics of concern, through topic modeling.

The dataset was first subset to the 60% where *Review_Sentiment* matches the *Star_Sentiment*, as this signifies that the user reviews contains text with sentiments that matches the user sentiments towards the app and would be more suitable for topic modeling. User reviews with neutral sentiments were omitted for deeper analysis, while reviews with positive and negative sentiments were then extracted separately for topic modeling. This was about 12.1% and 44.0% of the entire dataset respectively.

7.1 Model

Latent Dirichlet Allocation (LDA) was used for topic modeling. The *Tokenised_Review* dataset was first used to create the dictionary, and also converted to a bag-of-words format where each review consisted of a list of (token id, token count) tuples. There were two main LDA algorithms available, the online Variational Bayes algorithm by Hoffman, Blei, and Bach (2010), and the MAchine Learning for LanguagE Toolkit (MALLET)'s LDA wrapper which is based on Gibbs Sampling. The latter was selected for topic modeling, after a preliminary examination found that it gave more well-defined topics. The number of topics would be assessed based on the coherence score, which measures the quality of the learned topics, and the models' inter-topic distances.

7.2 Number of Topics

Topic modeling was conducted multiple times with differing number of topics (2 to 12). The coherence scores obtained (Figures 9 and 10) showed that 6-7 topics for negative sentiments, and 3-5 topics for positive sentiments, provided a good quality of learned topics.

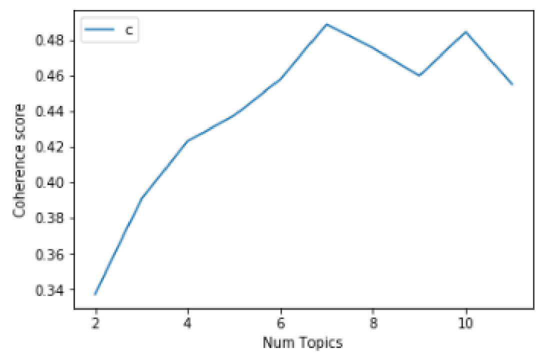


Figure 9: Coherence scores for MALLET's LDA, on reviews with negative sentiments.

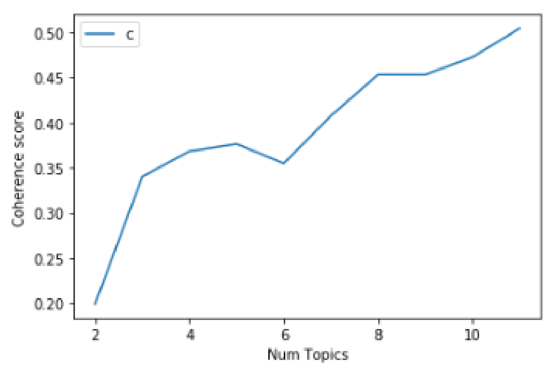


Figure 10: Coherence scores for MALLET's LDA, on reviews with positive sentiments

After assessing the models' inter-topic distances (Appendix B) and the quality of the topics amongst the varying topic numbers, we selected 6 topics for reviews with negative sentiments, and 3 topics for reviews with positive sentiments.

7.3 Topics

7.3.1 Negative Topics

The 6 main topics in negative reviews, through assessing the words obtained for each topic (Figure 11), were:

1. App malfunction: App crashing after an update, unable to launch app
2. Speed and authentication: slow loading speed, fingerprint authentication
3. Transaction issues: issues with transactions like fund transfers, card payment
4. Bad user experience
5. Digital token issues
6. Login issues: issues with password; password error; had to reinstall app, and login issues after an update.

```
[(0,
  '0.177*"update" + 0.077*"work" + 0.072*"fix" + 0.054*"unable" + '
  '0.051*"version" + 0.044*"crash" + 0.041*"open" + 0.033*"late" + '
  '0.024*"upgrade" + 0.019*"launch"'),
 (1,
  '0.116*"log" + 0.096*"login" + 0.071*"time" + 0.035*"page" + 0.035*"slow" + '
  '0.033*"load" + 0.032*"screen" + 0.031*"hang" + 0.026*"fingerprint" + '
  '0.021*"button"'),
 (2,
  '0.056*"make" + 0.056*"account" + 0.042*"transaction" + 0.032*"show" + '
  '0.025*"useless" + 0.025*"pay" + 0.025*"check" + 0.025*"payment" + '
  '0.022*"card" + 0.020*"enter"'),
 (3,
  '0.050*"bad" + 0.047*"access" + 0.043*"user" + 0.028*"function" + '
  '0.022*"call" + 0.021*"service" + 0.020*"experience" + 0.020*"force" + '
  '0.016*"good" + 0.016*"feature"'),
 (4,
  '0.080*"token" + 0.049*"transfer" + 0.042*"time" + 0.042*"work" + '
  '0.028*"digital" + 0.024*"sms" + 0.024*"option" + 0.022*"set" + 0.018*"fail" '
  '+ 0.018*"money"'),
 (5,
  '0.052*"app" + 0.042*"issue" + 0.041*"problem" + 0.034*"password" + '
  '0.033*"error" + 0.029*"change" + 0.026*"give" + 0.025*"reinstall" + '
  '0.025*"key" + 0.023*"pin"')]
```

Figure 11: Words for each negative topic.

7.3.2 Positive Topics

The 3 main topics in positive reviews, through assessing the words obtained for each topic (Figure 12), were:

1. App improvements and updates: Improvements in service and app functions after update
2. Interface simplicity and ease of use: Simple and easy-to-use app interface and features
3. User friendly experience and functionalities: Nice experience with account service and transactions

```
[(0,
  '0.079*"work" + 0.049*"update" + 0.029*"version" + 0.025*"improvement" + '
  '0.025*"improve" + 0.023*"previous" + 0.022*"login" + 0.022*"time" + '
  '0.016*"issue" + 0.015*"fix"'),
 (1,
  '0.171*"easy" + 0.045*"feature" + 0.039*"interface" + 0.036*"simple" + '
  '0.034*"make" + 0.030*"fast" + 0.018*"transfer" + 0.018*"function" + '
  '0.015*"pay" + 0.013*"navigate"'),
 (2,
  '0.075*"nice" + 0.070*"friendly" + 0.034*"experience" + 0.032*"convenient" + '
  '0.026*"job" + 0.026*"account" + 0.025*"design" + 0.025*"service" + '
  '0.024*"transaction" + 0.019*"lot"')]
```

Figure 12: Words for each positive topic.

7.4 Topics in User Reviews

Each user review was then labelled based on its dominant topic (*Review_Topic)*, defined as the topic number that had the highest percentage contribution in that review.

7.4.1 Negative Topics

App malfunctions (23.7%) were the most common complaint among users, including among users of the 3 largest banks, DBS, OCBC, and UOB (Figure 13). The second most common negative *Review_Topic* was speed and authentication issues (18.9%).

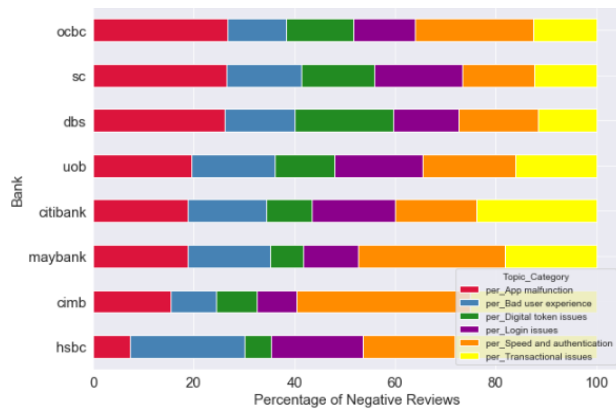


Figure 13: Proportion of Negative Reviews by *Review_Topic.*

7.4.2 Positive Topics

Overall, app improvements and updates (57.8%) were the most common compliment among users.

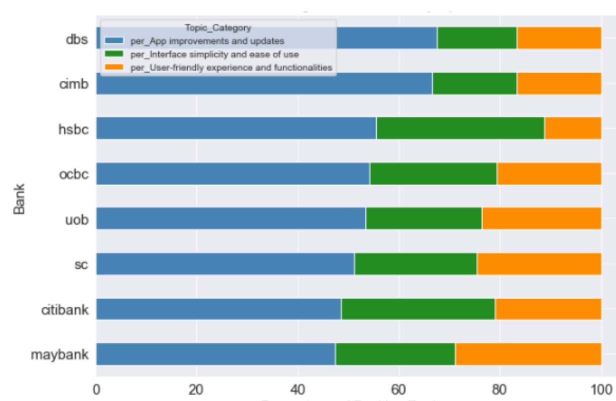


Figure 14: Proportion of Positive Reviews by *Review_Topic*

7.5 Results Interpretation

7.5.1 INSIGHTS FROM TOPIC MODELING

Based on the main topics extracted, we make the following general recommendations to banking app developers:

1. App security, which was cited in earlier reports as a chief concern of users, was not evident from reviews clustering. This indicates that banking app users generally trust that the necessary security measures would have been implemented.
2. Login difficulties and long loading times emerged as pressing issues, consistent with the findings of earlier studies. However, app malfunctions still make up majority of negative reviews, suggesting that developers should devote more resources towards enhancing and maintaining app stability.
3. Transactional and digital token issues were also revealed to be the cause behind many of user complaints. These factors were not previously flagged in earlier studies as major user concerns. However, their emergence as topics in our present study should remind developers that these problems also require urgent attention.
4. Improvements in app service and updates generated much more positive reviews than the other 2 major topics (Interface and User-Friendliness). This is indicative that apart from fixing the pressing outstanding problems, developers should also focus efforts on extending and enhancing services on the apps, if they wish to boost app popularity. While we are not suggesting that interface design and user-friendly features are unimportant, we recognise that service improvements are much likelier to strengthen the reputations of beleaguered apps.

7.5.2 RATIO OF POSITIVE-TO-NEGATIVE REVIEWS (TOP 3 BANKS – DBS, OCBC AND UOB)

We took a closer examination into the reviews of the 3 largest banks by assets, DBS, OCBC, and UOB, for examples of actionable business insights that could be derived through topic modeling.

The ratio of positive to negative reviews could be a possible business metric of how well an app is doing. An analysis of this metric on a monthly basis from Jan 2016 to Feb 2020 (Figure 15) revealed that generally, the UOB had been performing worse over time, while DBS and OCBC seemed to have been performing better over time. Of note, DBS had more positive reviews than negative reviews around end-2019, which is most likely related to their new app updates implemented in October and September 2019.

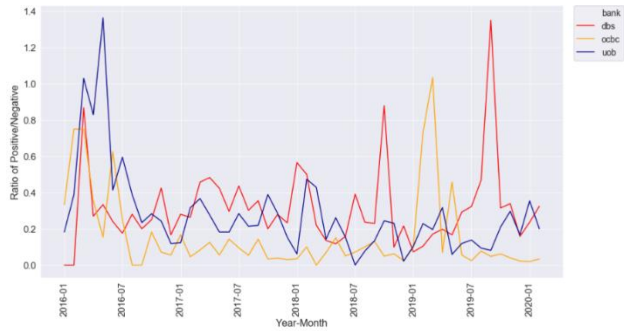


Figure 15: Monthly ratio of positive-to-negative reviews for DBS, OCBC, and UOB (Jan 2016 to Feb 2020).

7.5.3 EXAMINATION OF SPIKES IN REVIEWS (TOP 3 BANKS – DBS, OCBC AND UOB)

The number of positive and negative reviews for each topic could also provide business insights on areas that the apps are doing well. For instance, an analysis of the number of positive reviews due to app improvements and updates revealed that DBS had a spike in October 2019, again most likely due to their new app update. Based on both metrics, DBS should learn that the app update was welcomed by users, and thus perform a closer examination of the changes implemented and improve upon them.

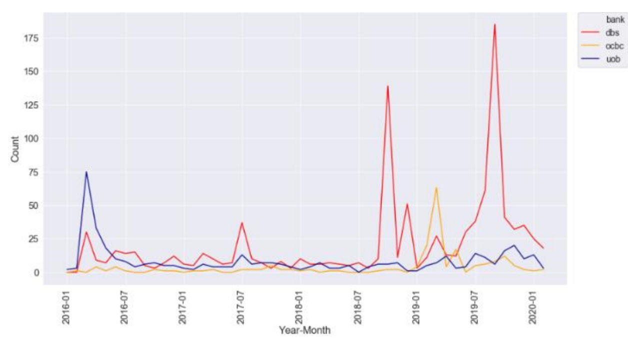


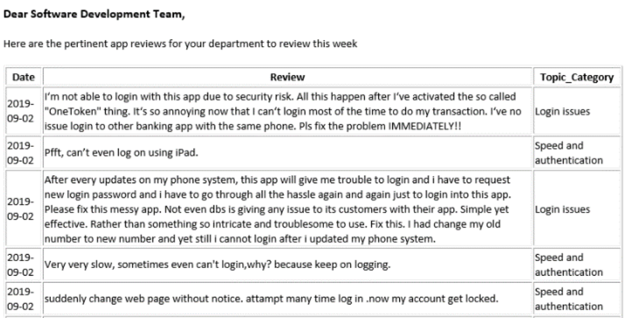
Figure 16: Monthly number of positive reviews due to app improvements and updates for DBS, OCBC, and UOB (Jan 2016 to Feb 2020).

The full set of graphs visualizing these topic trends for the top 3 banks can be found in Appendix C.

7.5.4 AUTO FORWARDING TO RELEVANT DEPARTMENTS

With user reviews segmented to different topics, banks could auto-forward reviews to relevant departments for their follow-up. For instance, reviews about digital token issues could be assigned to the app software development team, while reviews about bad user experience could be assigned to the UI/UX team and customer service team. An example of an email is in Figure 17 below.

This solution helps to create additional business value by ensuring that the complaints are correctly directed to the appropriate teams for investigation and quick implementation of necessary improvements.



Dear Software Development Team,

Here are the pertinent app reviews for your department to review this week

Date	Review	Topic_Category
2019-09-02	I’m not able to login with this app due to security risk. All this happen after I’ve activated the so called "OneToken" thing. It’s so annoying now that I can’t login most of the time to do my transaction. I’ve no issue login to other banking app with the same phone. Pls fix the problem IMMEDIATELY!!	Login issues
2019-09-02	Pfft, can’t even log on using iPad.	Speed and authentication
2019-09-02	After every updates on my phone system, this app will give me trouble to login and i have to request new login password and i have to go through all the hassle again and again just to login into this app. Please fix this messy app. Not even dbs is giving any issue to its customers with their app. Simple yet effective. Rather than something so intricate and troublesome to use. Fix this. I had change my old number to new number and yet still i cannot login after i updated my phone system.	Login issues
2019-09-02	Very very slow, sometimes even can't login,why? because keep on logging.	Speed and authentication
2019-09-02	suddenly change web page without notice. attampt many time log in .now my account get locked.	Speed and authentication

Figure 17: Example of email auto-forwarded to the app software development team.

The email auto-forwarding code was written in Python, and makes use of a SMTP (Simple Mail Transfer Protocol) server to forward messages to pre-specified email addresses of relevant departments. The message body is the HTML script that includes the data-frame filtered based on the date and *Review_Topic*. Depending on the bank’s requirements, the script can be modified to automatically collate and send reviews on a weekly basis.

8. Supervised Classification Models

As established in section 6.3.2, there is room to explore if supervised classification models can predict *Star_Sentiment* from the review text. In building the model, we first used 60% of the dataset where *Review_Sentiment* matches *Star_Sentiment*, where it was assumed that any text patterns of the review would match the *Star_Sentiment*.

It was further split to train and validation set based on 80:20 proportion. The *Star_Sentiment* was used as the data labels. The best models were then tested on the remaining 40% of the dataset where the *Review_Sentiment* did not match the *Star_Sentiment*.

8.1 Models

8.1.1 Multinomial Naive Bayes

The first model explored was the Multinomial Naive Bayes model, which assumes that every feature was independent of others. The model was fitted on the *Vectorised_Review* training dataset, with hyperparameter-tuning (alpha) and 5-fold cross-validation with GridSearchCV to reduce over-fitting. The best fit model resulted in a validation F1-score of 0.60 and a test F1-score of 0.42.

Table 7: Test classification report for Multinomial NB.

	Precision	Recall	F1-Score	Support
Negative	0.69	0.98	0.81	7,912
Neutral	0.02	0.00	0.00	1,479
Positive	0.81	0.31	0.44	3,261
Overall	0.51	0.43	0.42	12,652

8.1.2 Linear Support Vector Classification (SVC)

Given the large number of features, it was assumed that the dataset was likely to be well linearly separable. Therefore, we selected another linear classification model, the SVC with linear kernel. The model was similarly fitted on the *Vectorised_Review* training dataset, with hyperparameter-tuning (loss and C) and 5-fold cross-validation with GridSearchCV. The best model provided an improved validation F1-score of 0.67 and a test F1-score of 0.47.

Table 8: Test classification report for Linear SVC.

	Precision	Recall	F1-Score	Support
Negative	0.73	0.94	0.82	7,912
Neutral	0.07	0.03	0.04	1,479
Positive	0.74	0.43	0.55	3,261
Overall	0.51	0.47	0.47	12,652

8.1.3 CNN

We next explored a Convolutional Neural Network (CNN) model that considers the sequence of words. The model was based on three filter sizes to capture *n*-grams: 2, 3, and 4. A convolution layer and pooling layer is created for each filter size. This resulted in a model with eight hidden layers, including the embedding layer and a concatenation layer to merge the resultant feature maps. The *Padded_Review* dataset was used. The model resulted in a validation F1-score of 0.66 and test F1-score of 0.45.

Table 9: Test classification report for CNN.

	Precision	Recall	F1-Score	Support
Negative	0.73	0.86	0.79	7,912
Neutral	0.03	0.03	0.03	1,479
Positive	0.72	0.42	0.53	3,261
Overall	0.49	0.43	0.45	12,652

8.2 Model Comparison

A comparison of the models' F1-scores (Table 10) showed that the linear SVC algorithm was the best model in predicting the *Star_Sentiment* based on user reviews as it had the highest validation and test scores.

Table 10: F1-scores of supervised classification models.

Model	Validation	Test
Multinomial NB	0.60	0.42
Linear SVC	0.67	0.47
CNN	0.66	0.45

8.3 Results Analysis

The linear SVC model has a test accuracy score of 70%. As such, out of the 40% of reviews where *Review_Sentiment* do not match the *Star_Sentiment*, about 70% could be predicted by other textual features using the linear SVC model. This shows that about 10% *(40% x [100% - 70%] ≈ 10%)* of overall user reviews cannot be used to predict the *Star_Sentiment*. These reviews could be spam reviews, reviews containing sarcasm or cases where the users had rated the app wrongly. Banks could consider ignoring these reviews when analysing the performance of their apps.

9. Future Steps

Moving forward, we believe that there are several steps we can take in the future to improve the results of our analysis.

At the moment, the study is based on the number of reviews obtained from both app stores, without any data available on the total number of downloads. With the total number of downloads known, we will then be able to gain a better understanding of the app performances by looking at proportions of negative and positive reviews to total downloads, instead of looking at absolute numbers of reviews. This means that we can consider using metrics such as number of negative reviews per 1,000 downloads or number positive reviews per 1,000 downloads.

We also noticed that some of the reviews were in fact spam, where some reviews were of the exact same template, while others contained content clearly meant for game apps. Applying a spam filter will thus help to improve the results of the business insights delivered, as well as performances of the sentiment analysis and topic modeling.

Lastly, ensemble learning methods such as max-voting can potentially also improve the accuracy of the sentiments tagged to each review.

10. Conclusion

With the push to improve Singapore's financial technology ecosystem and the need to attract and retain customers in this increasingly digitalised world, banks should perform deeper analysis of their user reviews to assess areas where they have done well, and where they could do better.

References

Au Yong, J. (2016, April 14). MAS 'must balance innovation and trust in fintech push'. *The Straits Times.* Retrieved from https://www.straitstimes.com/singapore/mas-must-balance-innovation-and-trust-in-fintech-push

Badenhausen, K. (2019, March 4). The World's Best Banks: ING and Citibank Lead The Way. *Forbes.* Retrieved from https://www.forbes.com/sites/kurtbadenhausen/2019/03/04/the-worlds-best-banks-ing-and-citibank-lead-the-way/#661bd8565314

Cash is still king in Singapore as reluctant elders shun mobile apps. (2019, March 10). *Singapore Business Review.* Retrieved from https://sbr.com.sg/financial-services/in-focus/cash-still-king-in-singapore-reluctant-elders-shun-mobile-apps

Choo, Y. T. (2020, January 7). 21 applications submitted for up to 5 Singapore digital bank licenses: MAS. *The Straits Times.* Retrieved from https://www.straitstimes.com/business/banking/21-applications-submitted-for-5-singapore-digital-bank-licences

De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research, 13, 2031-2035.*

Hoffman, M. D., Blei, D. M., & Bach, F. (2010, December). *Online Learning for Latent Dirichlet Allocation*. 23rd International Conference on Neural Information Processing Systems 2010, Vancouver, British Columbia, Canada.

Hutto, C. J., & Gilbert, E. (2014, June). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14), Michigan: Ann Arbor.

Klein, N., Marinescu, I., Chamberlain, A., & Smart, M. (2018, March 6). Online Reviews are Biased. Here's How to Fix Them. *Harvard Business Review.* Retrievd from https://hbr.org/2018/03/online-reviews-are-biased-heres-how-to-fix-them

Lak, P. & Turetken, O. (2014, January). *Star Ratings versus Sentiment Analysis – A Comparison of Explicit and Implicit Measures of Opinions.* 47th Hawaii International Conference on System Science, Waikoloa, HI, USA. doi: 10.1109/HICSS.2014.106

Lam, F. (2019, June 17). Singapore's 3 big banks can defend market share against fintech threats: Moody's. *Business Times.* Retrieved from https://www.businesstimes.com.sg/companies-markets/singapores-3-big-banks-can-defend-market-share-against-fintech-threats-moodys

Pang, B. & Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization With Respect to Rating Scales. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 115-124). Ann Arbor: Michigan: Association for Computational Linguistics. doi: 10.3115/1219840.1219855

Ng, R. J. (2019, August 1). 2 in 3 Singapore bank customers keen to try digital banks: poll. *Business Times.* Retrieved from https://www.businesstimes.com.sg/banking-finance/2-in-3-singapore-bank-customers-keen-to-try-digital-banks-poll

Singaporeans ditch bank visits to go mobile. (2018, July 25). *Singapore Business Review.* Retrieved from https://sbr.com.sg/financial-services/news/singaporeans-ditch-bank-visits-go-mobile

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631-1642). Seattle, Washington, USA: Association for Computational Linguistics.

Tan, J. (2017, April 19). Singaporeans most demanding when it comes to good service. *Business Times.* Retrieved from https://www.businesstimes.com.sg/consumer/singaporeans-most-demanding-when-it-comes-to-good-service

Weisstein, F. L., Song, L., & Zhu, Y. (2017) Examining Impacts of Negative Reviews and Purchase Goals on Consumer Purchase Decision. *Journal of Retailing and Consumer Services, 39,* 201-207. doi: 10.1016/j.jretconser.2017.08.015

Why only 9% of Singaporeans prefer to use banking apps. (2017, July 11). *Singapore Business Review.* Retrieved from https://sbr.com.sg/financial-services/in-focus/why-only-9-singaporeans-prefer-use-banking-apps

Appendix A – Word Clouds

The word clouds for OCBC (Figures 1 and 2) revealed that customers are unhappy with security issues such as One-Time Password (OTP) and token, and happy that the app is user-friendly/easy to use. The word clouds for UOB (Figures 3 and 4) revealed that customers are unhappy with loading issues, and happy that the app is user-friendly/easy to use.

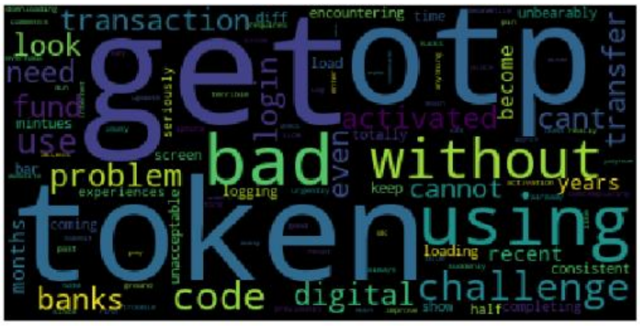


Figure 1: Word cloud of all negative reviews for OCBC.

Figure 2: Word cloud of all positive reviews for OCBC.

Figure 3: Word cloud of all negative reviews for UOB.

Figure 4: Word cloud of all positive reviews for UOB.

Appendix B – Visualisation of Inter-Topic Distances

The inter-topic distances for reviews with positive sentiments are in Figures 1 to 4; reviews with negative sentiments are in Figures 5 to 8.

Positive Reviews

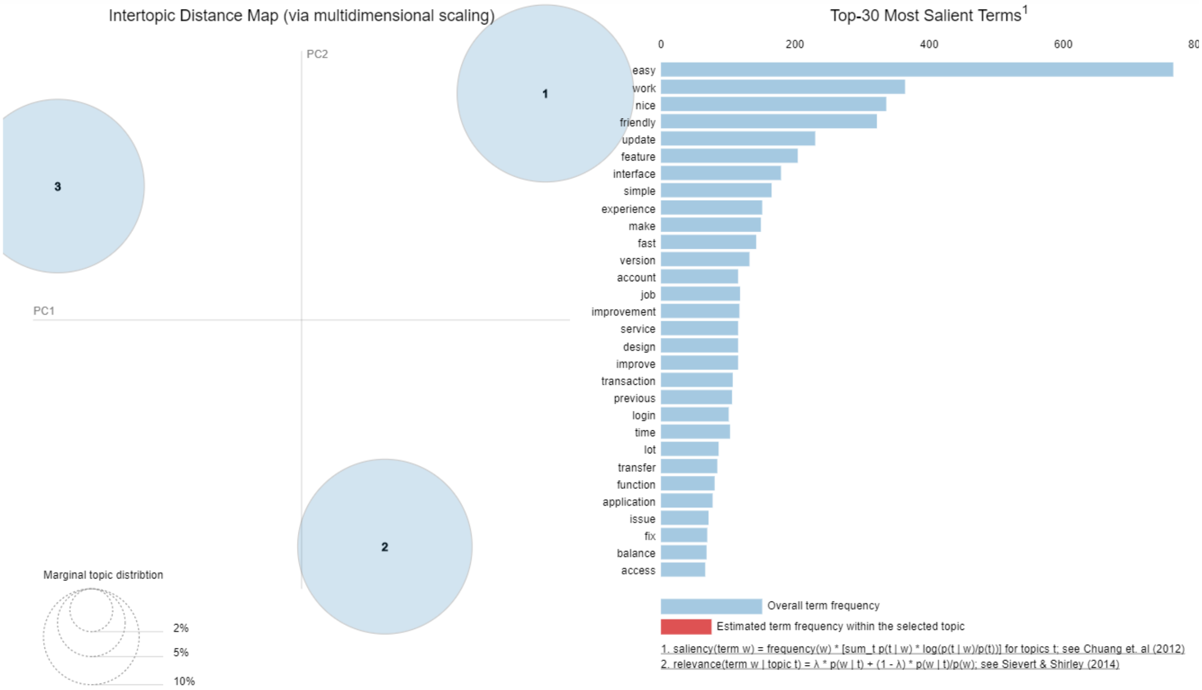


Figure 1: Positive reviews, 3 topics

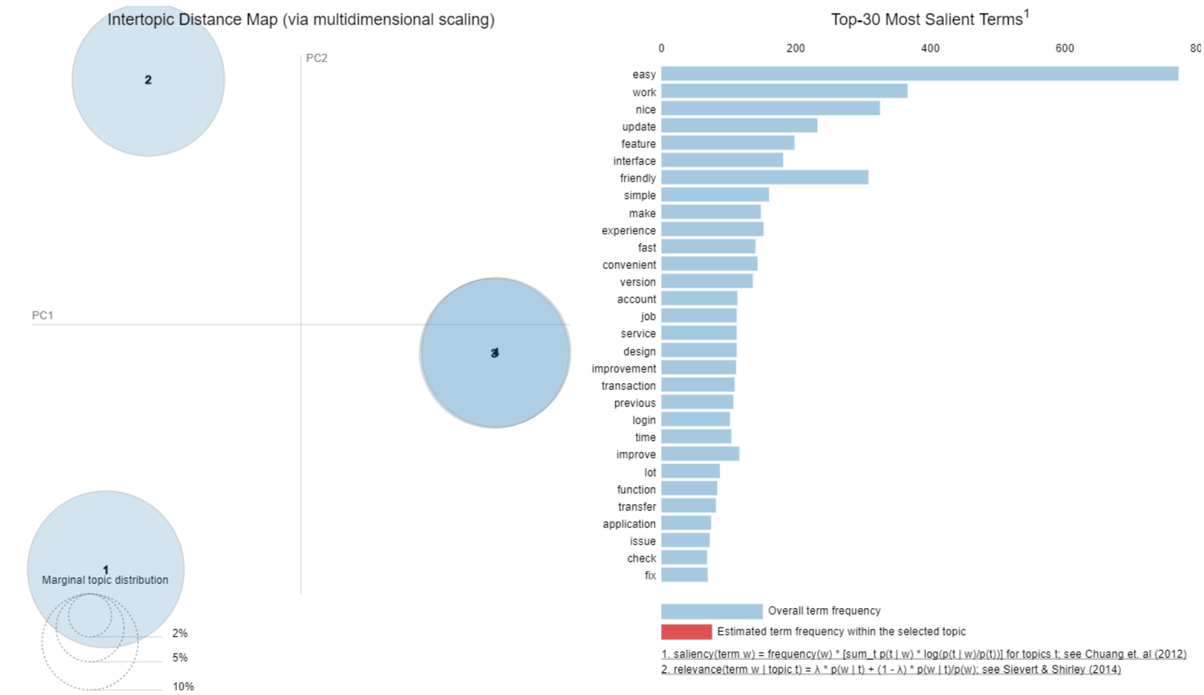


Figure 2: Positive reviews, 4 topics

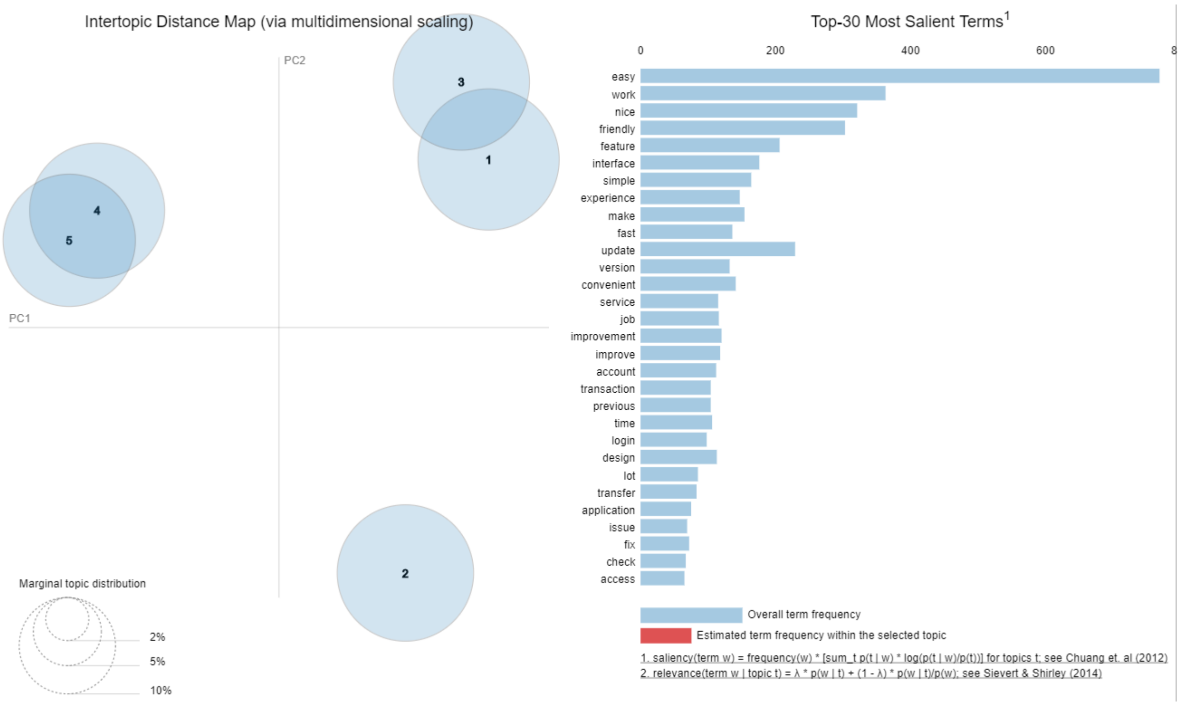


Figure 3: Positive reviews, 5 topics

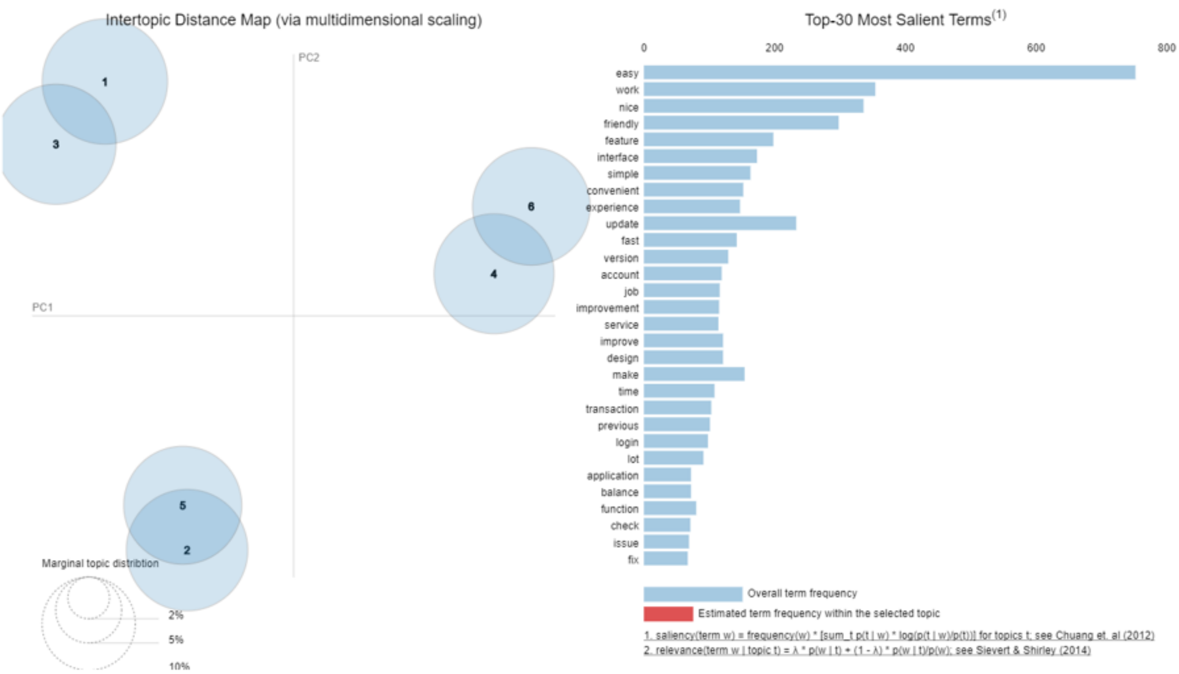


Figure 4: Positive reviews, 6 topics

Negative Reviews

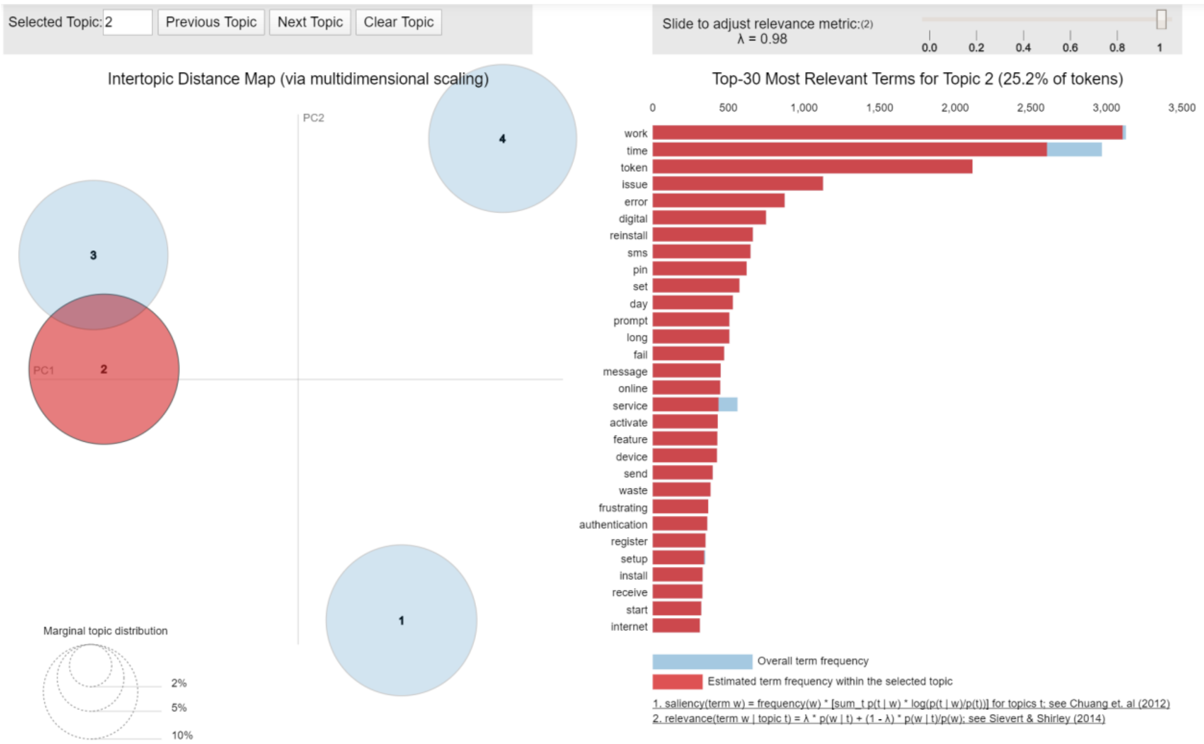


Figure 5: Negative reviews, 4 topics

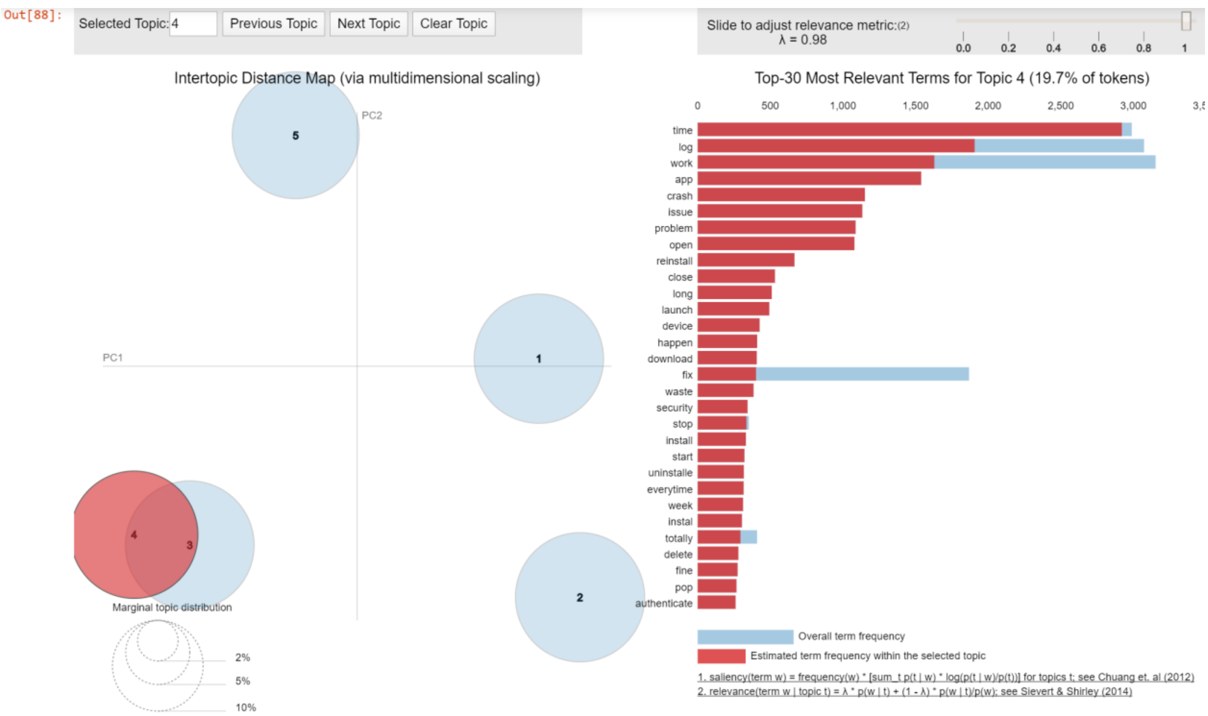


Figure 6: Negative reviews, 5 topics

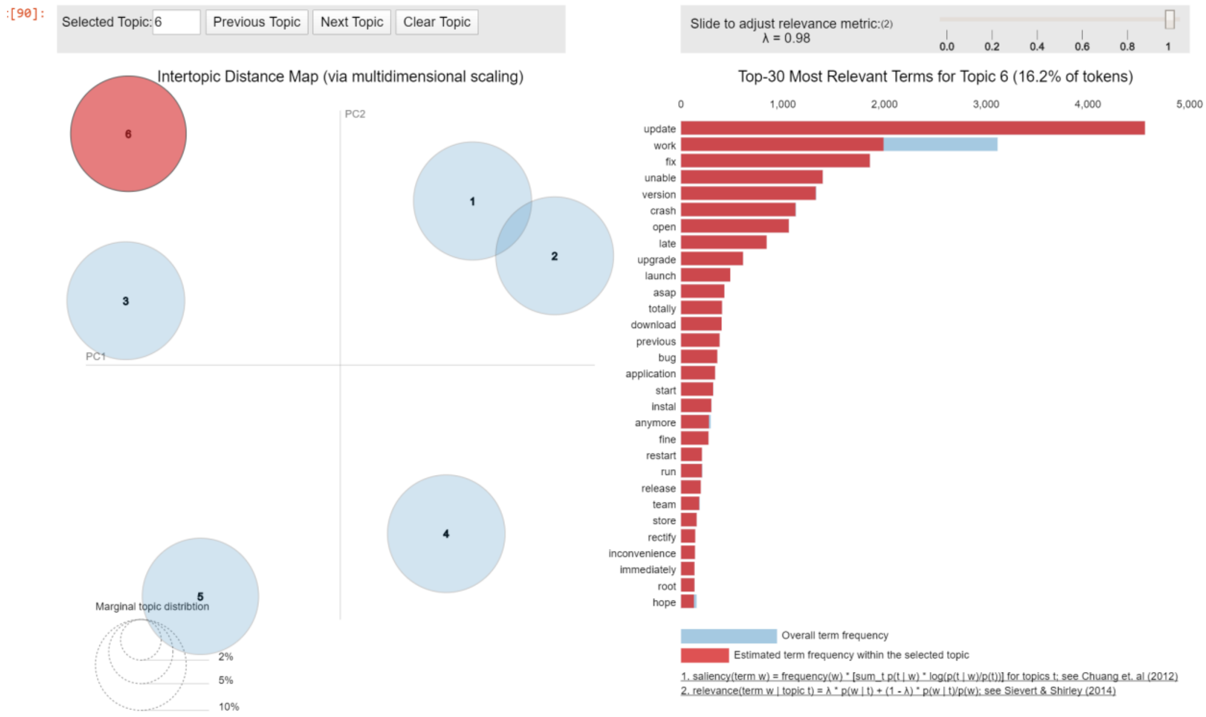


Figure 7: Negative reviews, 6 topics

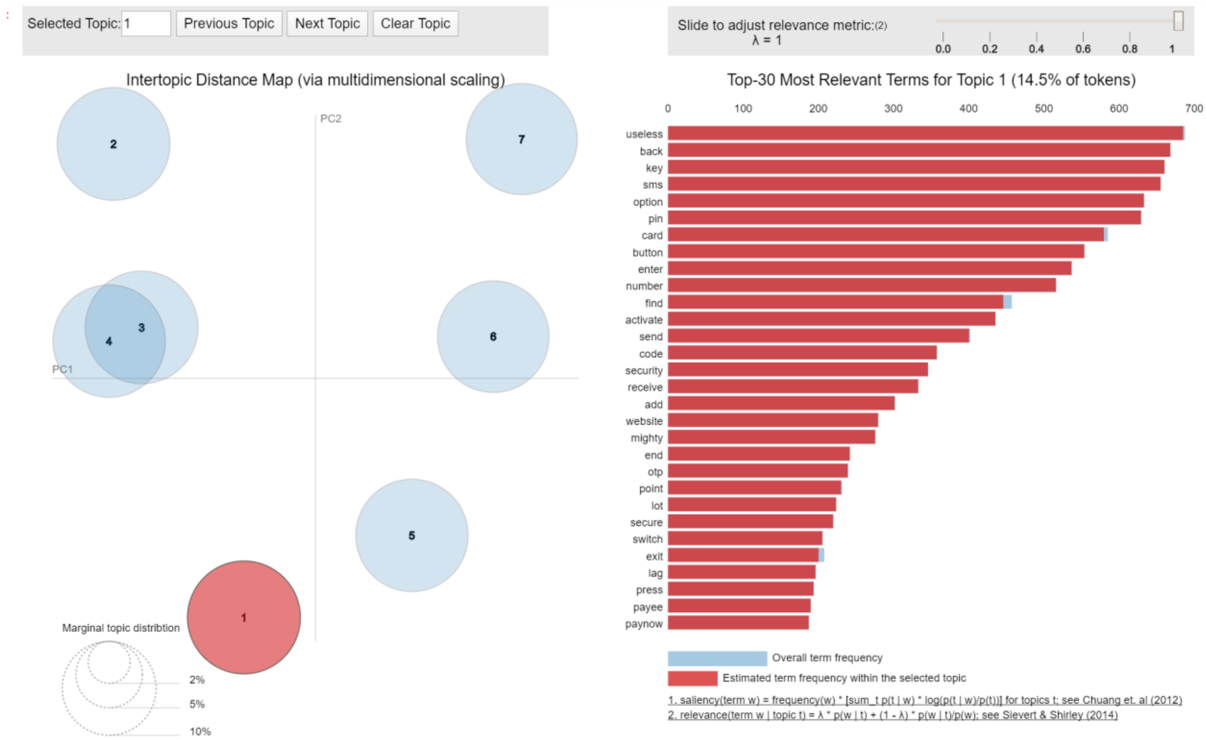


Figure 8: Negative reviews, 7 topics

Appendix C – Topic Trends for Top 3 Banks (DBS, UOB, OCBC)

The detailed trends for positive sentiment topics and negative sentiment topics for the apps of the top 3 banks (DBS, UOB, OCBC) are shown here.

Topics Trends of Positive Reviews

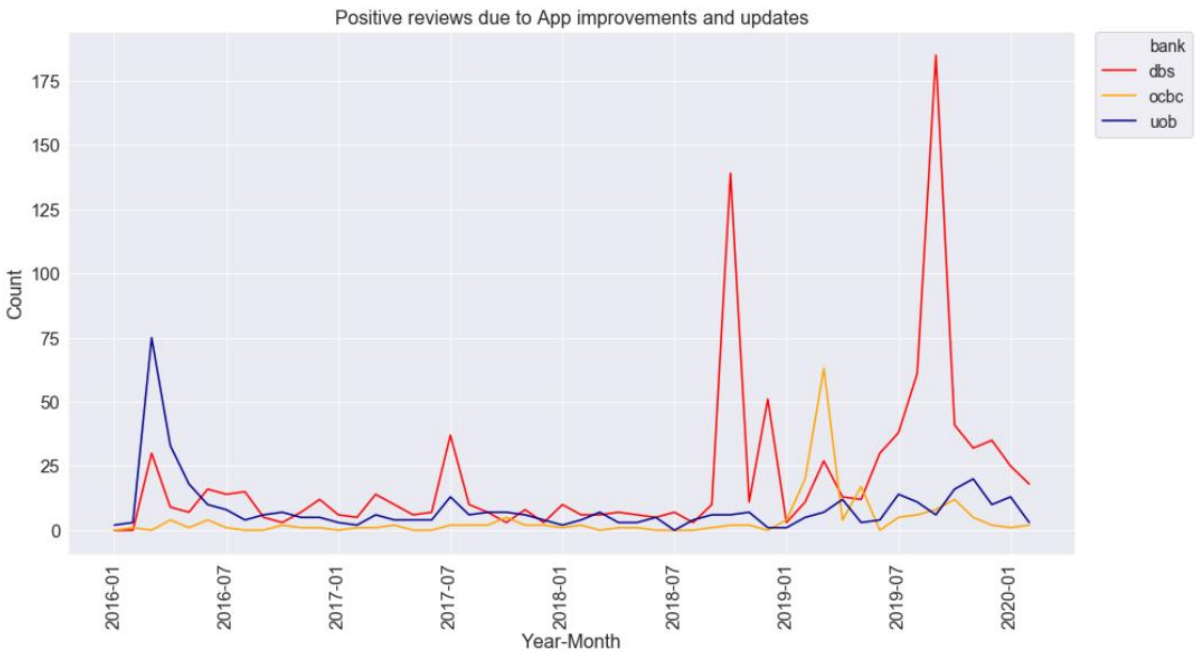


Figure 1: Trends of positive reviews relating to app improvements and updates

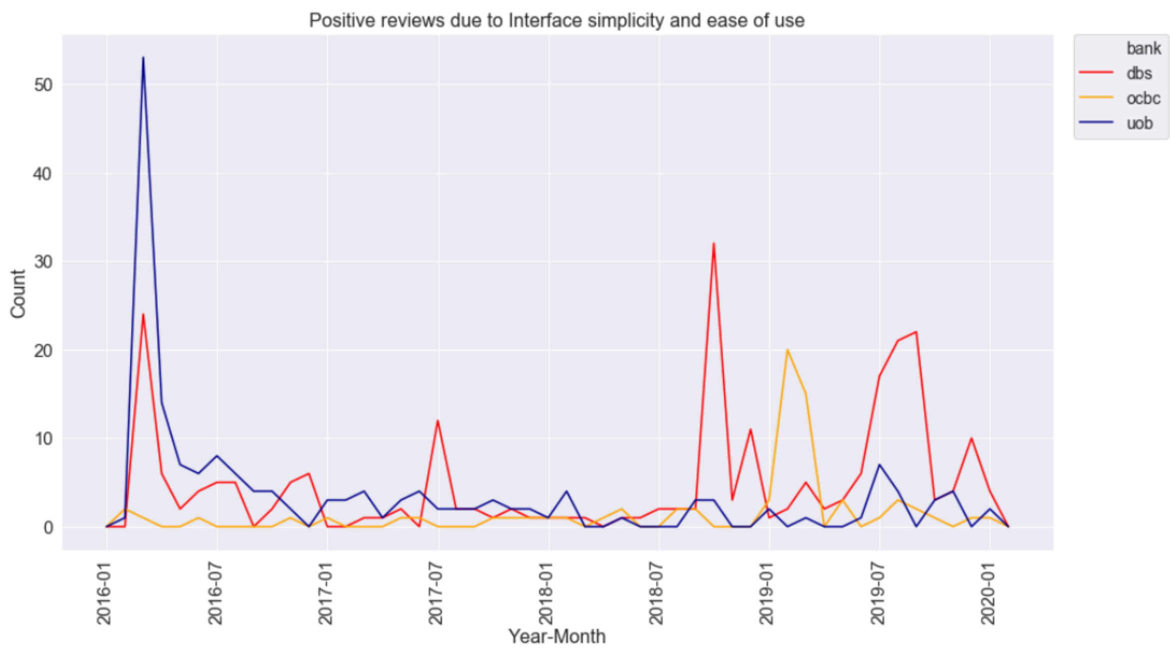


Figure 2: Trends of positive reviews relating to interface simplicity and ease of use

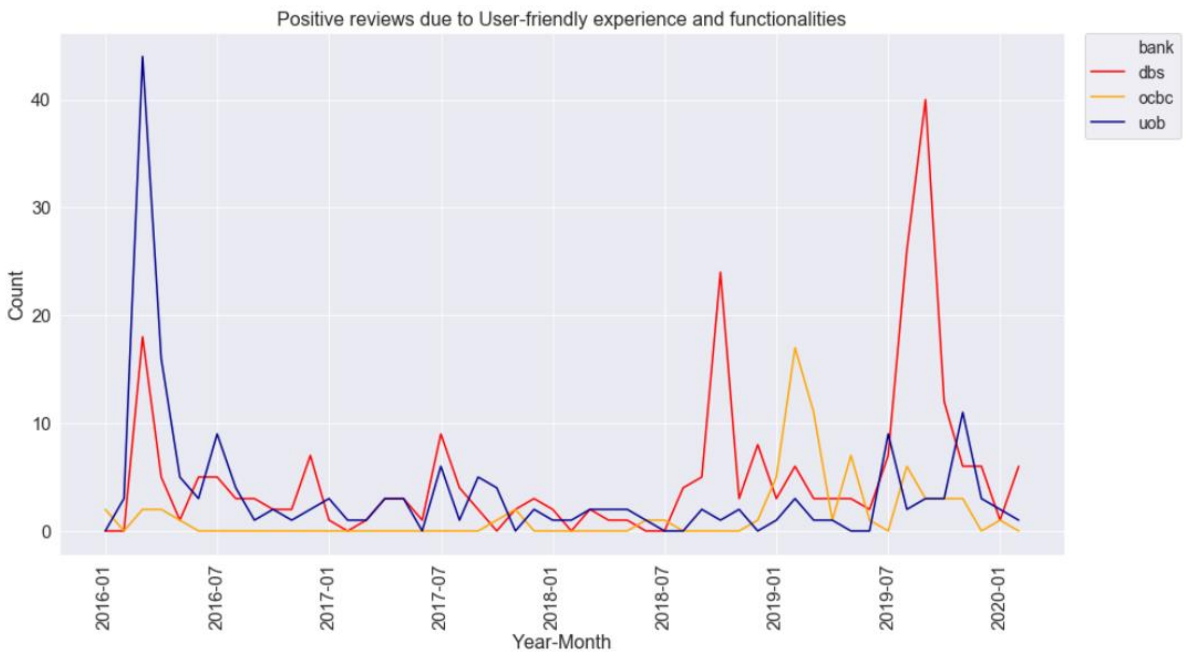


Figure 3: Trends of positive reviews relating to user-friendly experience and functionalities

Topics of Negative Reviews

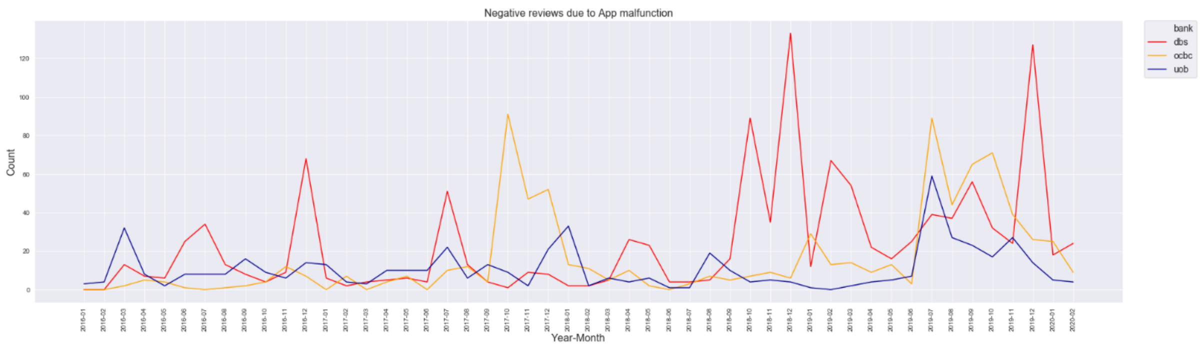


Figure 4: Trends of negative reviews relating to user-friendly experience and functionalities

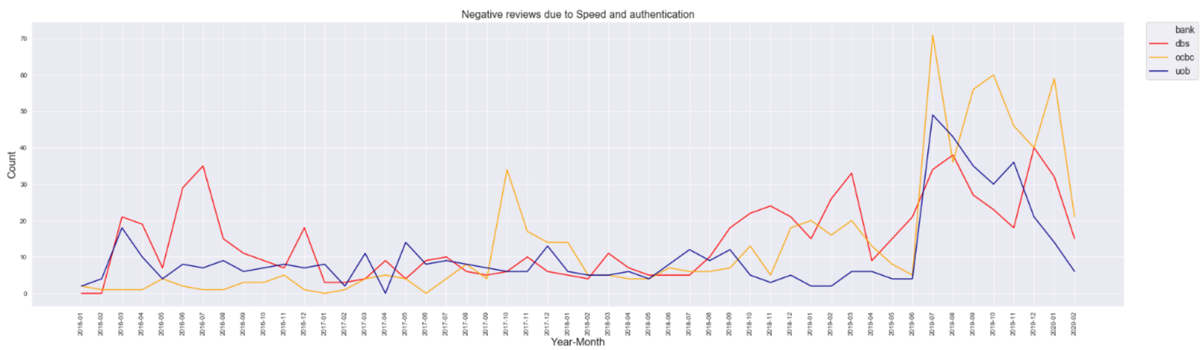


Figure 5: Trends of negative reviews relating to speed and authentication

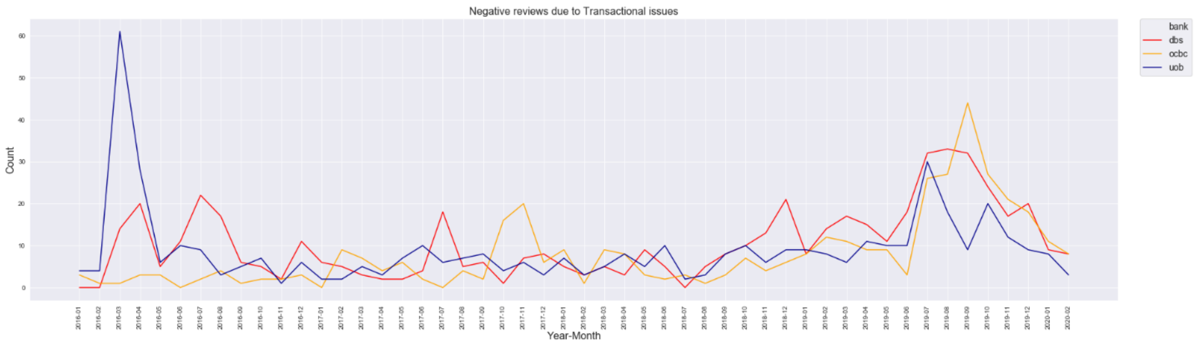


Figure 6: Trends of negative reviews relating to transactional issues

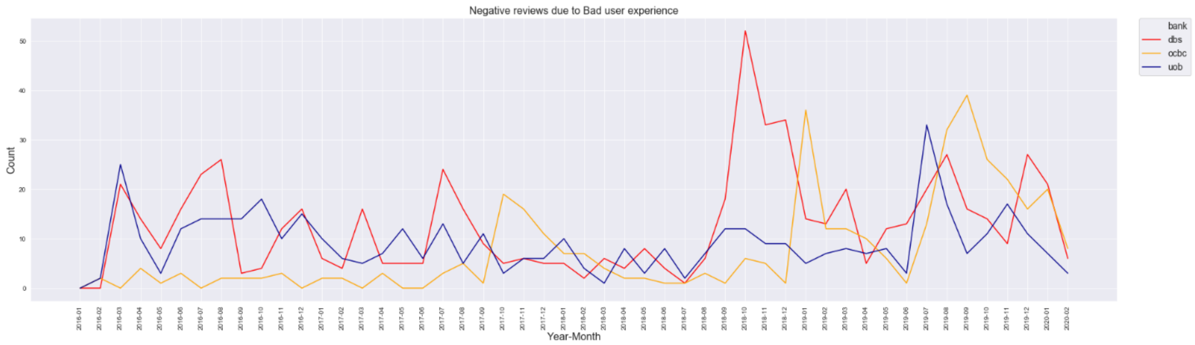


Figure 6: Trends of negative reviews relating to bad user experience

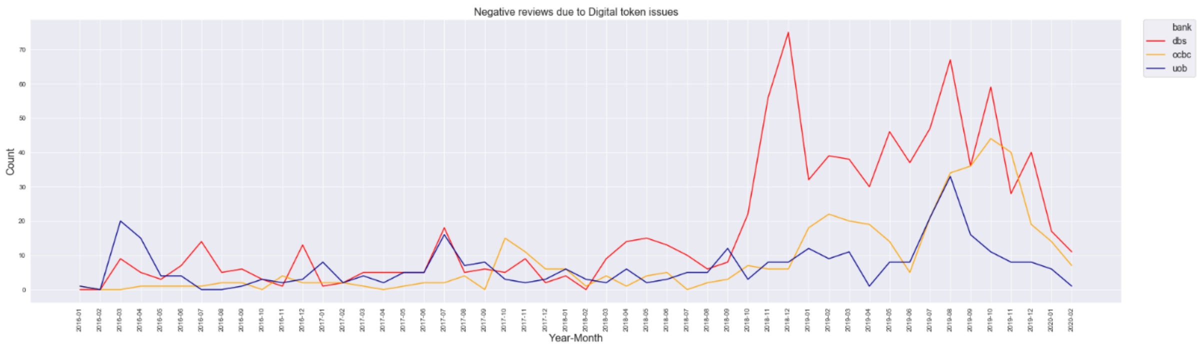


Figure 7: Trends of negative reviews relating to digital token issues

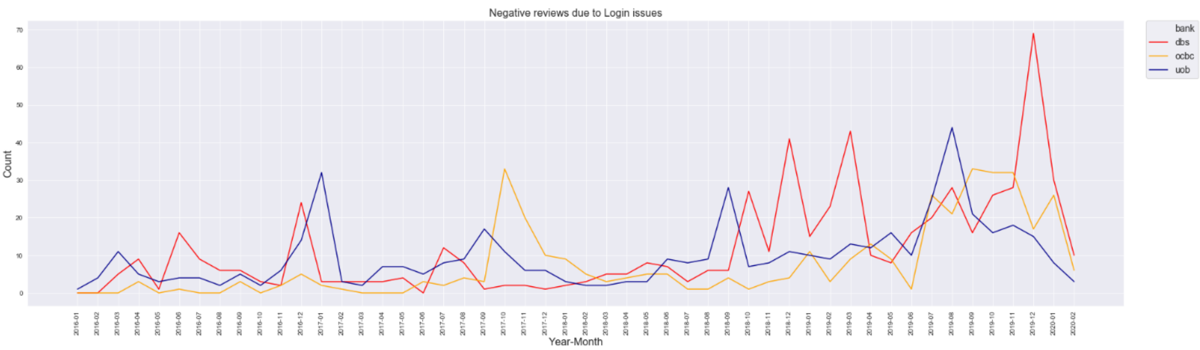


Figure 8: Trends of negative reviews relating to login issues