# SOCIAL MEDIA MONITORING TO ASSESS CONSUMER RISKS IN DIGITAL CREDIT APPS:

## Guidance for Supervisors from an India Pilot

CGAP

July 2023

Eric Duflos, Juan Carlos Izaguirre, Leena Datwani, Abhishek Mishra, Daryl Collins, Pravarakya Reddy Bathula, Leona Mathias, and Dhanashri Sawant

# ACKNOWLEDGEMENTS

This work was commissioned by CGAP and is the result of a collaboration with Decodis and the Reserve Bank Innovation Hub (RBIH).

The research team was composed of the following persons: Eric Duflos, Juan Carlos Izaguirre, Leena Datwani (CGAP); Daryl Collins, Pravarakya Reddy Bathula, Leona Mathias, Dhanashri Sawant (Decodis); and Abhishek Mishra (RBIH). Thanks to peer reviewers Silvia Baur-Yazbeck and Ivo Jenik (CGAP) and Aute Kasdorp (World Bank). This deck is part of the Social Media Monitoring implementation guide of CGAP's Market Monitoring for Financial Consumer Protection toolkit.

## Consultative Group to Assist the Poor

1818 H Street NW, MSN F3K-306
Washington DC 20433
Internet: www.cgap.org
Email: cgap@worldbank.org
Telephone: +1 202 473 9594


Cover slide photo credit: Sudipto Das via Communication for Development Ltd.

## Rights and Permissions

# TABLE OF CONTENTS

# I. BACKGROUND AND CONTEXT

Photo by Avishek Das

# What led up to this guidance?

In 2020, CGAP conducted a study in three countries, including India, to understand the effect of debt moratoria on low-income borrowers. By speaking to various stakeholders and analyzing social media data, we became aware of emerging consumer risks related with digital credit in India, in particular data misuse associated with irresponsible debt collection and data protection practices. In partnership with Dvara Research, in 2021 CGAP co-organized a roundtable with financial service providers, academics, researchers, industry, and consumer associations based in India to better understand the situation. A key takeaway from the roundtable was the need to better understand the digital consumer credit market in India and the risks they pose for customers.

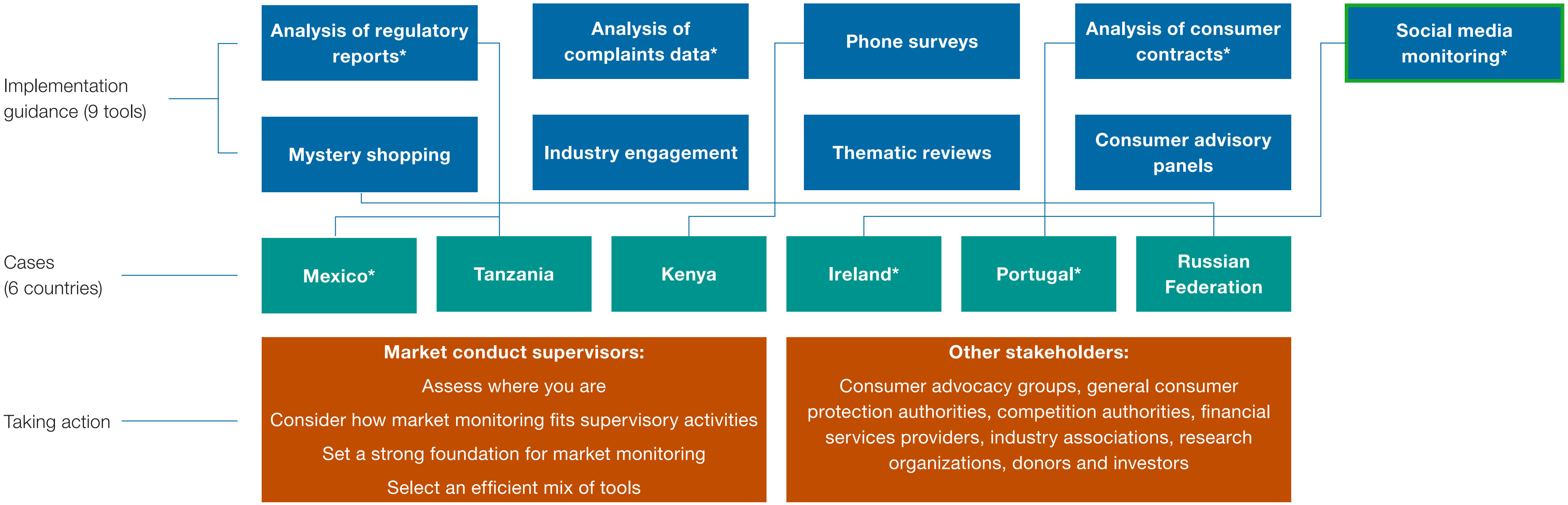
As a result, CGAP decided to pilot test a social media analysis tool based on Artificial Intelligence (AI), that can help to monitor the digital credit market, assess consumer risks, listen to the collective voice of consumers, and identify potentially concerning providers. The pilot used CGAP's new typology of digital finance consumer risks and CGAP's market monitoring toolkit (especially social media monitoring) as key frameworks. To conduct this pilot, CGAP contracted Decodis, a social research company, and collaborated with the Reserve Bank Innovation Hub (RBIH) of India.

This reading deck contains supervisory guidance on the use of a branch of AI, Natural Language Processing (NLP), for social media monitoring, based on insights and lessons from the India pilot, and provides examples of social media analyses carried out as part of that pilot. Although guidance is presented in a general manner that goes beyond the India pilot, it still needs to be well contextualized before its application in a specific jurisdiction.

Photo by Subrata Adhikary

# This guidance is part of CGAP's Market Monitoring Toolkit

CGAP's Market Monitoring Toolkit consists of a set of tools to enable authorities to carry out more preemptive and forward-looking consumer protection supervision. It aims to help supervisors identify, understand, and track financial consumer risks, behaviors, and outcomes by providing guidance to implement each tool, illustrative country cases, advice on how to act on market monitoring, and complementary resources. The implementation guidance for each tool lays out its benefits and opportunities, characteristics, how the tool can be used, limitations, and other useful resources. This deck is an integral part of the Social Media Monitoring implementation guidance.




* Guidance and cases that indicate the use of Suptech

Further information is available in Market Monitoring for Financial Consumer Protection

# AI-powered analysis of social media is relevant for monitoring consumer risks in digital lending apps

## 1. A new, ever-changing and potentially alarming digital lending app market is on the rise

The digital consumer lending app market is agile and fast-changing: Apps can be introduced into and taken off a store very quickly. In such a fast-changing market, new ways of assessing risks need to leverage data that are also available on a high-frequency basis. Digital lending apps are typically not regulated or supervised, and authorities cannot directly collect data on this market. Problems with digital borrowing can spill over to other markets.

## 2. Natural Language Processing has the power to detect complaint trends in social media

The complexity of text within social media data requires an AI tool that wades through mountains of text and finds trends. That is where NLP programming comes in. NLP benefits from being fast and automatable, which means that once established, it can run quickly. NLP skills have grown within call-center environments and enable proactive data analytics.

## 3. AI supports a customer-centric approach by listening to the collective consumer voice

Consumers may not easily access redress mechanisms for new apps but may directly raise their voices through social media. AI can then help listen to those voices and gather insights from consumer queries and complaints.

## 4. AI enables supervisors to identify practices from informal digital finance providers

Even if supervisors cannot ask for information from unregulated providers, they can use AI to indirectly monitor their practices through the analysis of social media data.

Further information on social media monitoring, customer-centric consumer protection, and collective consumer voice is available in Market Monitoring Toolkit – Social Media Monitoring, Elevating the Collective Consumer Voice in Financial Regulation, and Making Consumer Protection Regulation More Customer-Centric.

# How the social media analysis tool was piloted in India's digital lending app market

- The pilot used NLP – an important branch of AI and a key supervisory tool – Suptech – to analyze the social media data, identify **consumer complaints** related to various consumer risks and identify their degree of **urgency**.
- The pilot extracted data from Twitter and Google Play reviews relevant to digital lending apps in India for a first round of analysis in 2021 and a second round in 2022 to further iterate and test the NLP tool.
- The pilot used the number of downloads of these digital lending apps to estimate the number of users to **"size" the extent of the problem**. However, the pilot faced challenges by not being able to identify user characteristics as well as users of multiple apps.
- Lastly, the pilot used this data to identify **apps that warrant further attention**.

## Important

The objective of this pilot was to demonstrate what can be found through NLP analysis of social media content and to test and validate the tool with a second round of analysis.

**Making this tool fully ready for live supervision would require developing and testing it over time in the market where it will be used, considering conditions in the country context, such as the prevalence of different social media channels.**

# Complaints were categorized according to CGAP's typology of digital financial services consumer risks

CGAP identified 66 consumer risks in digital financial services, and categorized them into four broad risk types and two cross-cutting risks:

**Four broad risk types**

| FRAUD | DATA MISUSE | LACK OF TRANSPARENCY | INADEQUATE REDRESS MECHANISMS |
|---|---|---|---|
| Examples:<br>SIM swap fraud<br>Mobile app fraud | Examples:<br>Algorithmic bias<br>Unfair practices e.g., social shaming | Examples:<br>Undisclosed fees<br>Complex user interface | Examples:<br>Complex redress process<br>Expensive complaints handling system |

**Two cross-cutting risk types**

**AGENT-RELATED RISKS**
Examples: Liquidity challenges, agent fraud, discrimination based on social status

**NETWORK DOWNTIME**
Examples: Distributed Denial of Service (DDoS) attacks, insufficiently tested system upgrade, power outage

Fraud and data misuse are directly linked to cybersecurity. The two cross-cutting risks share some elements with all four broad risk types.

Further information is available in The Evolution of the Nature and Scale of DFS Consumer Risks.

# II. KEY LESSONS LEARNED

Photo by Vijay Pandey

# What this monitoring tool does effectively…

- ## The tool can track the nature of consumer risks in digital financial services

This approach analyzes daily or weekly data with very agile and automated methods, with market tracking on a daily or weekly basis. This market monitoring tool enables a nearly real-time understanding of developing trends of consumer risks in the markets, including whether levels of urgency are increasing.

- ## It tells us something about the consumers who are experiencing those risks

The code can detect different types of vulnerability among consumers, including those who are serial borrowers.

- ## It can detect early warning signals in digital financial services, especially emerging ones

Ongoing, high-frequency data can pick up growing numbers of complaints quickly, thereby picking up signals of problems well before the analysis of less frequent regulatory reporting data can be undertaken. This is especially important for emerging digital financial services and providers that may not yet be regulated or supervised.

- ## It can identify apps that merit their inclusion in a watch list

This approach is based on high-frequency data and analysis can be automated, detecting and tracking large and potentially concerning apps early. The tool helped detect several apps that were on the 2021 watchlist pilot. They were removed from the Google Play Store but when they were relisted in 2022, the tool detected them once again.

- ## Once coding is established, it can run inexpensively and frequently

Setting up the coding to collect, clean and analyze social media data is the most substantial investment, but thereafter, the expense of running this tool is negligible.

For more information on the overall benefits and opportunities of social media monitoring, see the Social Media Monitoring implementation guidance of CGAP's Market Monitoring Toolkit.

# …And recognizing its limitations

- ## The tool cannot detect the number and nature of complaints from those who cannot use social media

  Not all digital lending users have the agency, knowledge, or tools to put their complaints on any social media channel, like posting a “tweet” or completing a Google Play review. They are most likely to be the most vulnerable users.

- ## It cannot size the complaints in the market

  It is extremely difficult to detect how many people are using any sort of financial app because not all downloaders actually use the apps and therefore, not all users write reviews or do ratings. Moreover, because the handles of users are hidden it is not possible to tell whether there are users who use multiple apps.

- ## It does not enable the disaggregation of complaints by gender or income

  It is nearly impossible to determine which social media posts are associated with which gender because individuals who post content can hide their identity. With respect to income, with these data, it is possible to estimate types of livelihoods but not level of income.

For more information on the overall limitations of using social media data sources for market monitoring, see the Social Media Monitoring implementation guidance.

# What a supervisor needs to implement this tool

## • A strong foundation for the effective use of market monitoring tools

This includes an adequate legal mandate to perform market conduct supervision (specifically market monitoring), adequate staff (including analytical and subject matter experts to work with a vendor), and capacity to ensure high-quality data (including data protection requirements vis-à-vis personally identifying information).

## • Skills from a third-party vendor

Most of the skills needed within a third-party resource are available in local markets and though the process for putting this tool in place is not insignificant, it will run easily and inexpensively as the code gets more sensitive in time.

## • Technical capacity to recruit and manage a specialized vendor

It is important that the arrangements with the vendor ensure that they train, support, and transfer knowledge to the authority for adequate long-term implementation of the tool.

## • An initial investment in resources at a market conduct authority

In addition to the resources needed to contract a vendor, the market conduct authority needs to invest in internal resources to ensure that relevant supervisory and supporting staff such as IT specialists are informed and knowledgeable of the steps to take, implement and run the tool on an ongoing basis, and ensure high-quality data – especially because the tool improves with use.

This tool can be pragmatic, cost-effective, and complementary to other market monitoring tools that are available for market conduct and consumer protection authorities. To learn about the criteria for selecting a specialized vendor, see page 18. For more information on how to build a strong foundation for market monitoring, visit Taking Action in the Market Monitoring Toolkit.

# A significant advantage of this tool: it improves with use

- ## Continuous usage improves the "intelligence" and cost-effectiveness of the tool

  As new phrases and wording are introduced, the coding becomes more sensitive and can therefore pick up more signals earlier on. These improvements over time increase the return on the initial investment for this tool.

- ## New categories of complaints, users or behavior can be added

  As new data are continuously added, new terms are detected. This means that the capabilities of this tool expand as time goes on. The coding for these new categories can be re-run on old data to learn more about trends over time.

- ## Ongoing exposure between tool implementers and other regulatory units creates a productive dialogue

  Discussions about how the tool detects the impact of key policy or industry actions can be built into the tool's capabilities. In this way, key issues from the tool analysis are shared with other units in clear formats and language.

For more information on how to use social media monitoring, see the Social Media Monitoring implementation guidance. For an illustrative case of continuous usage of social media monitoring, see the Central Bank of Ireland country case.

# Supervisory objectives this tool can help with

## Examples from the India pilot

### 1. Identifying market risks proactively

- Google Play reviews provided an “early warning signal” to highlight consumer risk in digital lending apps. The analysis picks up consumer complaints before they appear in traditional media, capturing attention and spreading more widely.
- Twitter showed “persistence” of complaints even after apps were removed from the Google Play Store, which shows that this analysis can rapidly assess the effects of industry actions.

### 2. Assessing the urgency and nature of consumer problems

- 25% of all combined complaints in Twitter posts and Google Play reviews are tagged as urgent.
- 45% of Twitter complaints and 16% of Google Play review complaints revealed that the nature of consumer problems concerned aggressive debt collection, which causes high stress among users.

### 3. Improving the understanding of consumer problems

- Claims of “fake apps”, for instance, were 24% of Twitter complaints and 29% of Google Play review complaints, i.e. one of the most common complaints. NLP allowed a better understanding of why consumers thought the apps were fake (e.g. there were high overlaps with claims of hidden fees and unresponsive complaint procedures).

### 4. Generating a watch list of apps that merit further investigation

- As of May 2022, there were 8 sizable digital lender apps that should have warranted further investigation, comprising 41% of the market.

# III. DEVELOPING AN IMPLEMENTATION ROADMAP


Photo by Yunaidi Joepoet

# Suggested steps

1. Identify supervisory objectives that this tool can help with
2. Select a vendor that can help with the implementation of the tool, including helping to facilitate the identification of supervisory objectives
3. Select social media platforms to find data on complaints from the consumers
4. Run a pilot to determine parameters to track (for example, types of complaints), which would help monitor the market
5. Create NLP code for determined parameters, including gathering the data and building code to tag them into different categories based on the supervisory objectives
6. Use and improve the tool for market monitoring.

The following sections provide details on each of these steps.

# i. Identifying supervisory objectives

Photo by Somrah Sahu

# Possible supervisory objectives this tool could be used for

1. Determine the nature of risks that consumers face using ANY type of digital financial services (not just digital lending)
2. Identify unscrupulous actors in the market
3. Monitor changes in complaints, especially increases
4. Identify changes in urgent complaints, especially increases
5. Identify an increase in serial borrowing or other concerning customer behaviors
6. Identify the nature and changes in vulnerability of customers
7. Monitor the uptake of digital financial services that are not yet regulated

Note: For the India pilot, the first round of data analysis focused on 1, 2, 3, 4, and 7. The second round addressed 5 and 6. Each authority should prioritize a set of supervisory objectives based on its context (e.g. market maturity, supervisory capacity). Objectives can be expanded upon with time.


# ii. Selecting a vendor

Photo by Ullas Kalappura

# What to look for when selecting IT vendors to help with tool implementation

## 1. Expertise in social media analysis and NLP

- The vendor should have footprint in social media analysis and be able to scrape the data from platforms such as Twitter, Facebook and Google Play Store.
- Individuals in the vendor team should have extensive experience in building machine-learning models in NLP and present insights using a visualization tool.
- Individuals in the vendor team should also have extensive Python experience.

## 2. Ability to implement API integration

- Running this analysis as a market monitoring tool will require API integration with third-party platforms such as MeaningCloud for topic classification.
- The scraping tool will need to be integrated.
- Individuals in the vendor team should have experience creating platforms with integrated third-party tools.

## 3. Ability to automate the process

- The crucial step is to automate the steps involved from scraping the data to analyzing them using NLP and providing a dashboard to visualize the important market changes and status.
- Individuals in the vendor team should have experience building reporting dashboards.
- The vendor should also provide support to supervisors, to help them understand the process and use the tool—during the pilot exercises and afterwards.

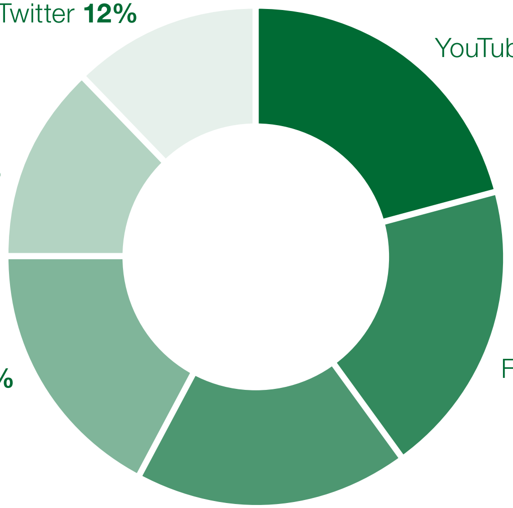
# iii. Selecting social media platforms

Photo by Jan Mohammad Shaikh

# How to choose which social media platforms to analyze

## 1. Consider the volume and share of users of different platforms in the country, and disregard those with too few users

For example, in India, there are 448 million adult active social media users, and all the main platforms have sizable shares of users (over 10%).




*Source: Statistica, as of January 2021, self-reported usage of each platform in the month.*

## 2. Evaluate what insights each key platforms can generate and how they can support market monitoring

**Google Play** reviews are open to the public and allow positive and negative reviews of apps, which can be relativized by the number of app downloads.

**Twitter** enables public discussion, advocacy and generating a collective voice.

**YouTube** is for sharing videos - including ads - but not for consumer views.

**Facebook** tends to be peer-to-peer and not about sharing views for the public to see (and the tool to scrape); apps have Facebook pages and answer questions via Facebook Messenger, but they focus more on advertising.

**WhatsApp** and **Instagram** have public groups, but usually by invitation or by receiving a sharable link.

## 3. Choose more than one platform to gather complementary data for market monitoring purposes

In India, the main need was to measure consumer complaints on emerging lending apps, so Google Play and Twitter best fit these needs.


# iv. Running a pilot to determine parameters to track

Photo by Vijay Pandey

# The parameters, or consumer complaint types, that can be tracked

| Consumer risk (per CGAP's typology) | What parameter will be tracked for each risk? | What does a social media post with this complaint look like? (Examples from the India pilot) |
|---|---|---|
| | TWITTER | |
| **Fraud** | **Information/Identity stealing**<br>ID information could be used for ill-intentioned purposes | This number 8792662456 from my cash app. Call me back to back and call my contacts without my permission. they calling my contacts and misusing my personal data. |
| | **Fake app or scam**<br>Complaints that apps deliberately fail to register payments or do not provide loans at all | #BanLoanApps<br><br>@RBI @RBIsays @CyberGujarat @DelhiPolice @GujaratPolice @BanegaAb @RBI @SCSC_Cyberabad<br><br>Loanfront Digital landings app scams m Already paid my loan amount but civil showing me not paid @ GoLoanFront |
| | GOOGLE PLAY | |
| **Data misuse** | **Inaccurate data**<br>Concerns about apps having/reporting wrong data about them | Please update correct consumers credit score details, many banks & nbfc's are giving wrong details to systems, especially HDFC BANK..... please check & verify details then update... it is humble request to your system please first collect all data from banks & nbfcs then update.... |
| | **Aggressive marketing or cross-selling**<br>Complaints of excessive calling pushing undesired services | This app ruined my credit score, after login so many marketing calls and sms received from various sources, it means paisa bazar leaks our important data, useless and senseless customer care Executives [...] |

# The parameters, or consumer complaint types, that can be tracked (continued)

| Consumer risk (per CGAP's typology) | What parameter will be tracked for each risk? | What does a social media post with this complaint look like? (Examples from the India pilot) |
|---|---|---|
| | **TWITTER** | |
| **Lack of transparency** | **Hidden terms**<br>Complaints that terms for loan repayment, disbursement, or interest rates were not fully disclosed at the time of contract<br><br>**Aggressive debt collection**<br>Complaints of harassment through visits, excessive SMS, or calls to customer and others in their contact list | "@Early_Salary your third party recovery agents call me 10 times today and also send me dhamki massage daily morning..and also you charge me 10000 rs extra intrest charges. please send me charges details #OperationHaftaVasooli@indSupremeCourt @KiritSomaiya https://t.co/yZ4ztYswBB"<br><br>'@zeebusiness @AnilSinghvi #OperationHaftaVasooli Sir kisi door ke relative ne credit card ka bill nahi pay kiya. They found me on Facebook nd started calling me on my mobile also on landline and using very bad abusive language. I don't know what to do.. |
| | **GOOGLE PLAY** | |
| **Inadequate redress mechanisms** | **Unresponsive complaints procedure**<br>Complaints about not being able to reach customer service or not getting a satisfactory response from them<br><br>**Complaints channels too costly**<br>Complaints of transactional costs or time-consuming process | I have repaid my loan amount but still app showing active loan.. emailed sent so many time no response received. Too slow customer service.<br><br>Worst first impression, It has been 13 days already the money is not yet disbursed to my Bank account where it's said they ONLY take 1 business day. No way to speak to any customer care service over a phone call and if u talk about emails it remains unanswered everytime... Every step is check marked except the last one where my money should be disbursed.... Now i fear that the loan is not disbursed and they are planning to auto debit EMI they have signed a mandate... This is ridiculous. |

# v. Creating NLP code for determined parameters

Photo by Vijay Pandey

# Steps to create NLP code for determined parameters

## 1. Data scraping process

The process of collecting text data from social media posts.

## 2. Data cleaning process

The process of removing some characters that will hinder the coding process.

## 3. Create topic modeling and topic classification codes for specific complaints

These are two different types of coding processes that analyze the data in different ways to detect types of complaints.

## 4. Create topic modeling and topic classification codes for urgency

These are two different types of coding processes that analyze the data in different ways to detect when complaints are urgent.

## 5. Create topic modeling and topic classification codes for vulnerable customers

These are two different types of coding processes that analyze the data in different ways to detect when those who are issuing complaints are vulnerable in different ways.

## 6. Create topic modeling and topic classification codes for serial borrowers

These are two different types of coding processes that analyze the data in different ways to detect when those who are issuing complaints have borrowed before.

Further information is available in the Technical Annex.


# vi. Using and improving the tool for market monitoring

Photo by Vijay Pandey

# Type of monitoring this tool does well and not so well

**Does exceptionally well**

## 1. Tracking complaints, urgency, vulnerability, and serial borrowing over time

Because this approach makes use of high-frequency data with very agile and automated methods, ongoing time series not only on a monthly but even weekly or daily basis are possible. Therefore, this market monitoring tool enables a nearly real-time understanding of developing trends of consumer experiences and risks in the markets. Moreover, any changes in market policy or providers' practices can be tracked to determine if they have an impact.

**Does exceptionally well**

## 2. Detecting a watch list of apps

Likewise, because this approach is based on high frequency data and analysis can be automated, detecting and tracking large and potentially concerning apps can be done earlier and any redressal actions taken against specific actors can be tracked to see if they are effective.

**Does not do well**

## 3. Sizing the complaints in the market

It is extremely difficult to detect how many people are using financial apps because not all downloaders actually use them and not all users write reviews or do ratings. Moreover, because the handles of users are hidden it is not possible to tell whether there are users who use multiple apps.

**Does not do well**

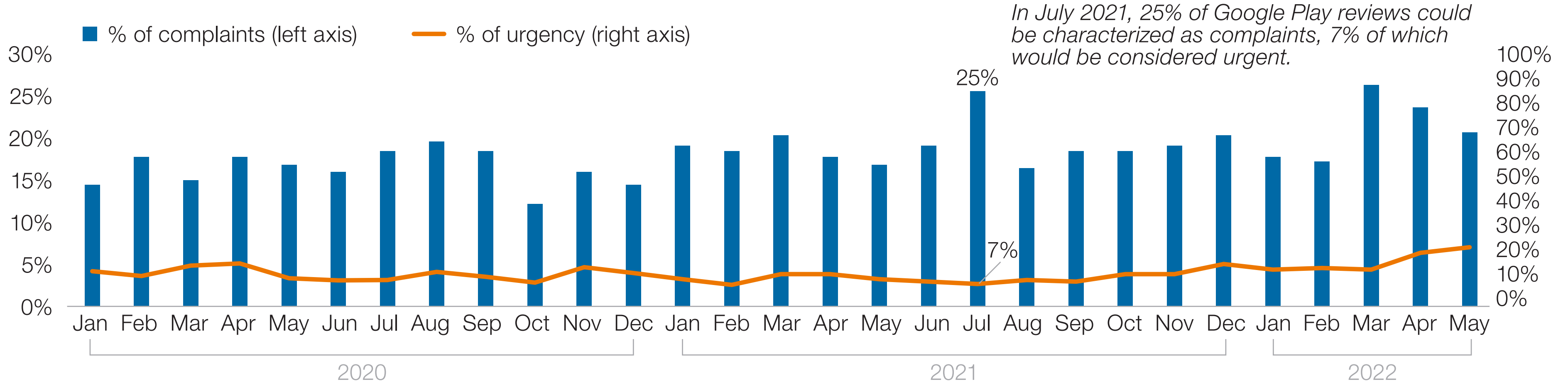
## 4. Detecting complaints by gender or income

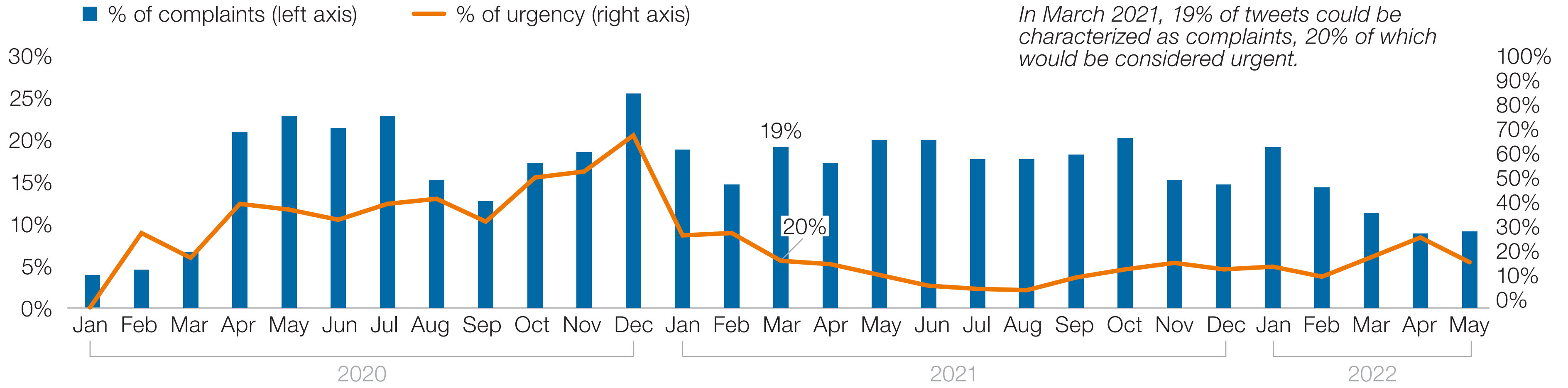
It is nearly impossible to determine which social media posts are associated with which gender because individuals who post content can hide their identity. Complaints never provide clues about income.

# India pilot: Tracking total and urgent complaints

## Share of posts denoted as complaints and share of complaints denoted as urgent, 2020-2022

The blue bar on the chart depicts the percentage of social media posts categorized as complaints every month and the orange line depicts the percentage of complaints labeled as urgent out of the total complaints for that month.



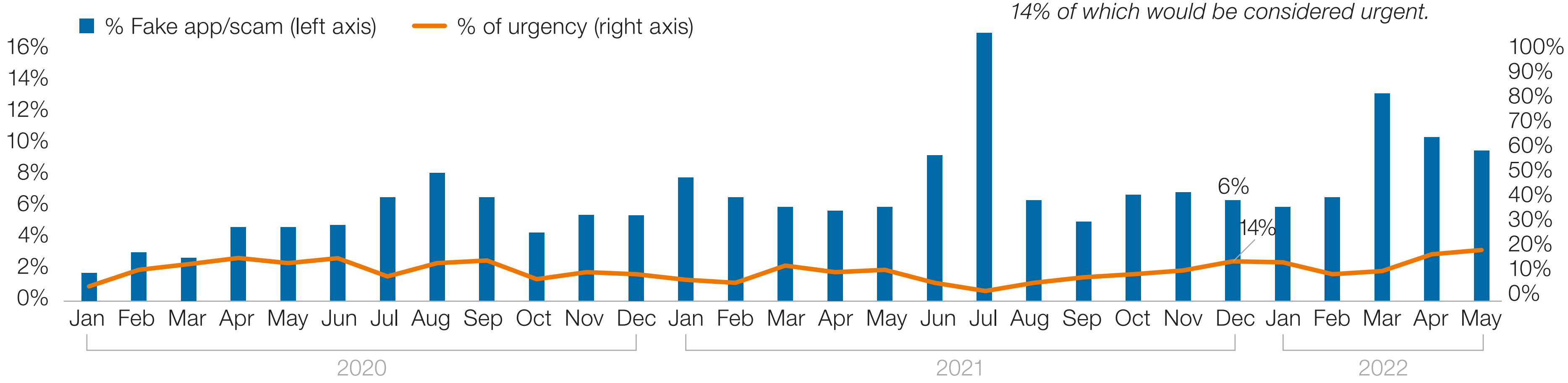



# India pilot: Tracking specific complaints per month

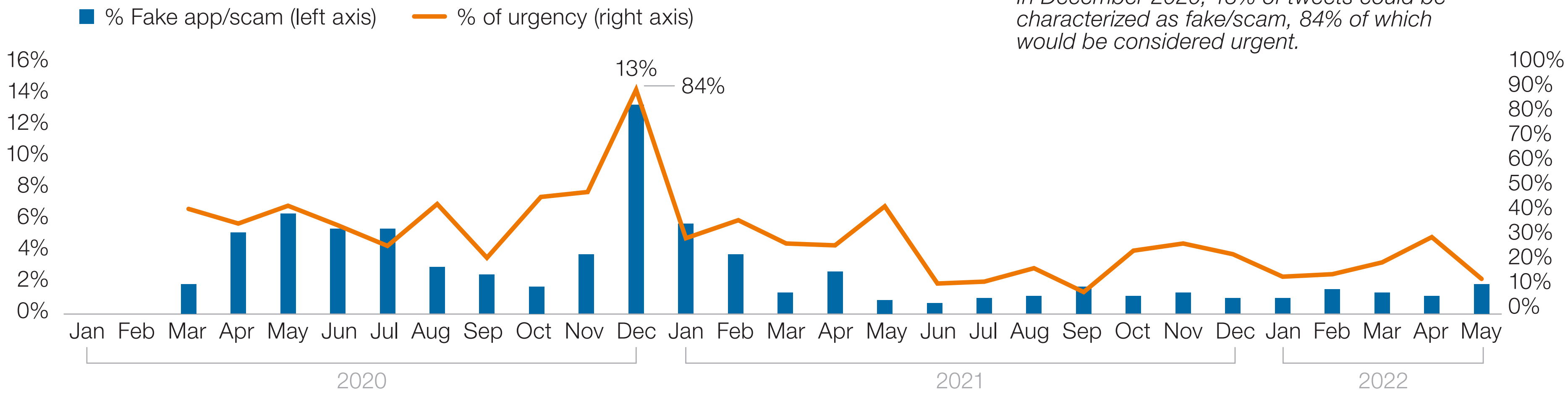
## Complaints about fraudulent apps (fake apps/scams), 2020-2022

The blue bar on the chart depicts the percentage of complaints on fraudulent apps as a percent of total complaints every month and the orange line depicts the percentage of complaints labelled as urgent out of the total fake app complaints for that month.

*In December 2021, 6% of Google Play reviews could be characterized as fake/scam, 14% of which would be considered urgent.*



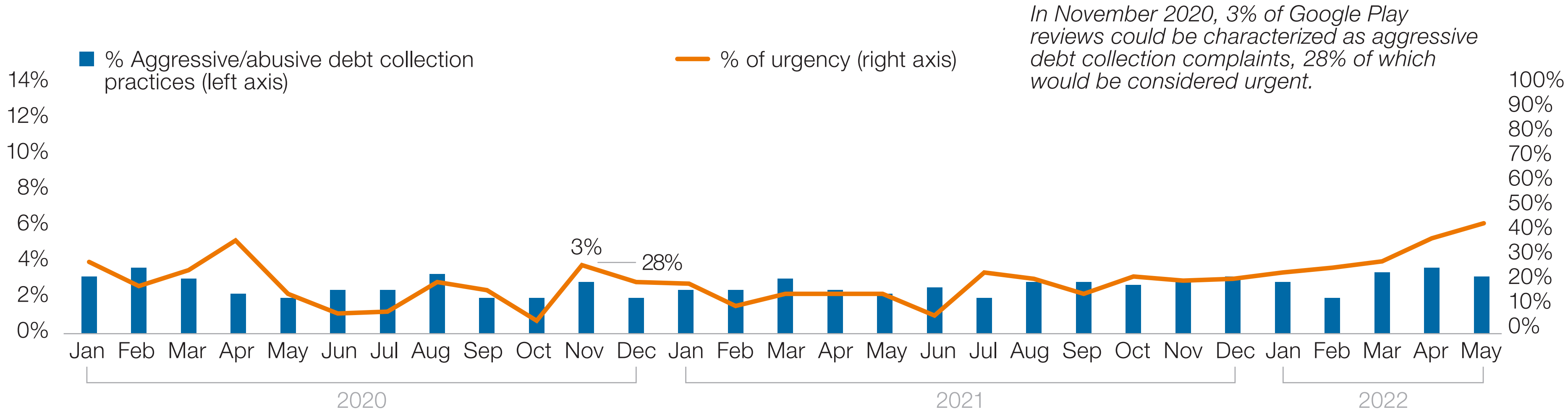

*In December 2020, 13% of tweets could be characterized as fake/scam, 84% of which would be considered urgent.*

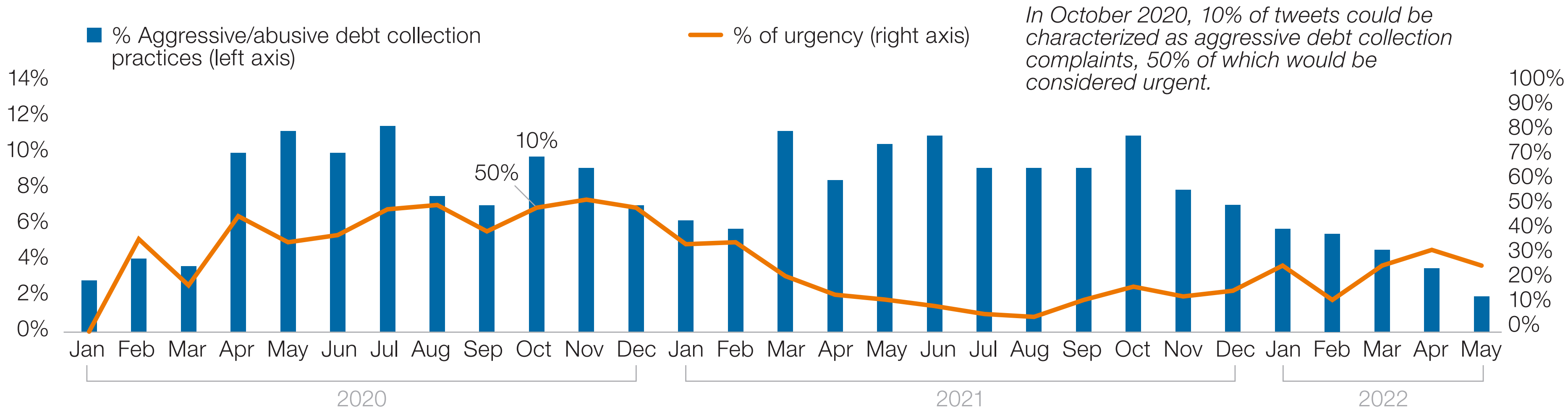


# India pilot: Tracking specific complaints per month

## Complaints about aggressive debt collection, 2020-2022

The blue bar on the chart depicts the percentage of complaints on aggressive debt collection as a percent of total complaints every month and the orange line depicts the percentage of complaints labelled as urgent out of the total aggressive debt collection complaints for that month.



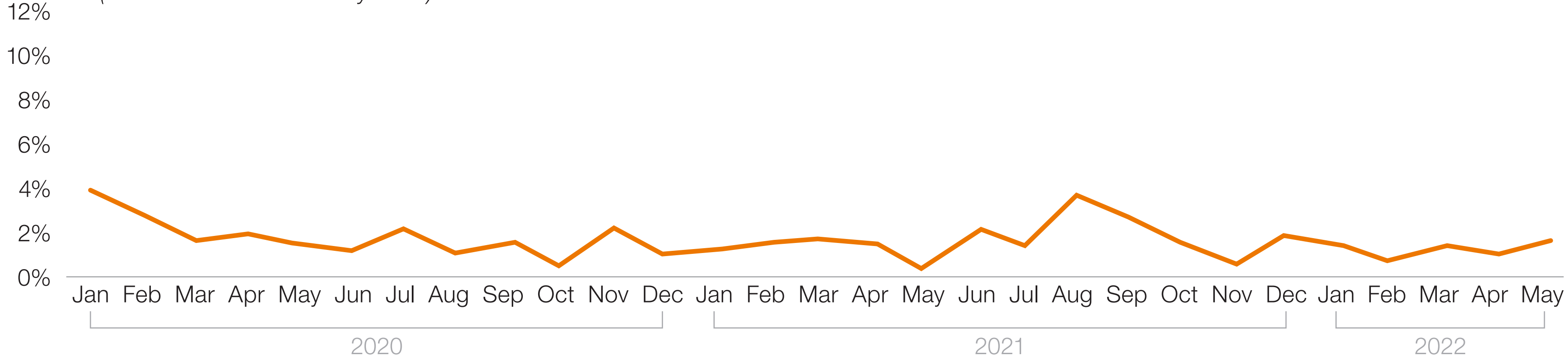
# India pilot: Tracking vulnerability[1]

## Complaints categorized as vulnerable, 2020-2022

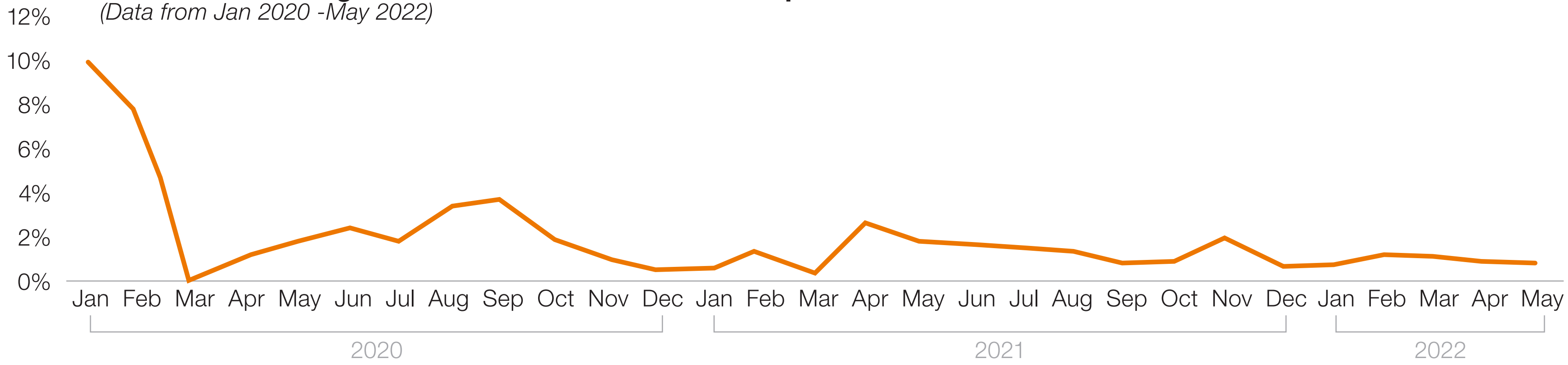
[1]Vulnerable = Lost job / Lack of work/ Not received payments; sickness; can't afford food/medicine; other as shown in slide 50.

The Unemployment rate in India between February 2022 and April 2022 was in the range of 7.6%-8.1%. Data showing the rise in unemployment post-January 2022 is available at CMEI.

**% of Google Play reviews categorized as vulnerable under complaints**
*(Data from Jan 2020 -May 2022)*

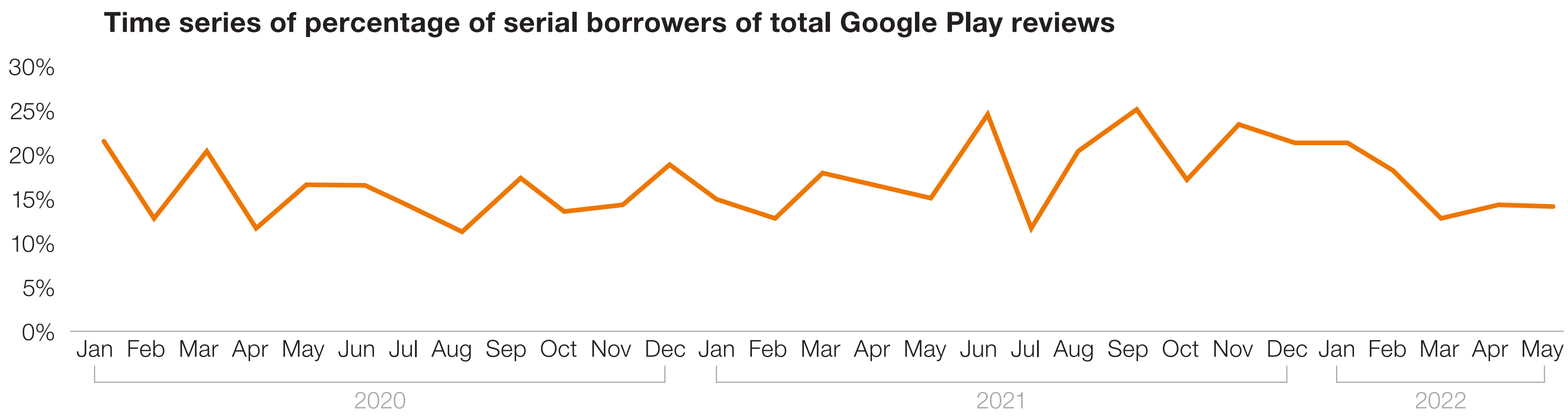



**% of Tweets categorized as vulnerable under complaints**
*(Data from Jan 2020 -May 2022)*



# India pilot: Tracking serial borrowers

## Back-to-back borrowing, 2020-2022



# India pilot: Identifying apps to watch

| Watchlist Jan 2020 – May 2021 | | | | |
|---|---|---|---|---|
| Ranking | App | % market share[1] | % urgent complaints | % aggressive debt collection complaints |
| 1 | Alpha | 6% | 21% | 49% |
| 2 | Beta | 6% | 28% | 40% |
| 3 | Charlie | 6% | 30% | 36% |
| 4 | Delta | 3% | 34% | 39% |
| 5 | Echo | 1% | 11% | 58% |
| 6 | Foxtrot | 6% | 39% | 24% |
| 7 | Golf | 6% | 15% | 35% |
| 8 | Hotel | 3% | 34% | 34% |
| 9 | Island | 1% | 32% | 43% |
| 10 | Juliet | 3% | 20% | 41% |
| 11 | Kilo | 6% | 24% | 24% |

Note that the table above reflects data from actual apps, but the names of those apps have been changed to conceal their identity. Note as well that this ranking has changed from the previous slide deck as Twitter data has now been included.

[1] Based on number of downloads from Google Play Store

**The watchlist** was calculated with a view towards balancing **market size and risk**.

While this tool was still in the pilot phase, weights were ascribed to the three indicators below to rank apps to watch. Changing the weights by +/- 10 basis points does not dramatically change the ranking.

| Indicator | Weight | |
|---|---|---|
| % market share[1] | 70 | |
| % urgent complaints | 10 | RISK |
| % aggressive debt collection complaints | 20 | RISK |

The pilot chose these indicators and assigned these weights because urgency and aggressive debt collection were important indicators during the COVID period. As the market changes, other indicators may become more important, and the watch list should be updated.

As more data accumulates through this tool, more sophisticated means, such as regression, can be used to develop the weights.

# India pilot: How the watchlist of apps changes over time

| Watchlist Jan 2020 – May 2021 | | | | |
|---|---|---|---|---|
| Ranking | App | % market share | % urgent complaints | % aggressive debt collection complaints |
| 1 | Alpha | 6% | 21% | 49% |
| 2 | Beta[1] | 6% | 28% | 40% |
| 3 | Charlie | 6% | 30% | 36% |
| 4 | Delta | 3% | 34% | 39% |
| 5 | Echo | 1% | 11% | 58% |
| 6 | Foxtrot | 6% | 39% | 24% |
| 7 | Golf | 6% | 15% | 35% |
| 8 | Hotel | 3% | 34% | 34% |
| 9 | Island | 1% | 32% | 43% |
| 10 | Juliet | 3% | 20% | 41% |
| 11 | Kilo | 6% | 24% | 24% |

11 most concerning apps = 47% of market

[1] Beta is CashBean. In February 2022, the Reserve Bank of India canceled the registration certificate of CashBean's parent company. See https://www.rbi.org.in/scripts/BS_PressReleaseDisplay.aspx?prid=53317.

| Watchlist May 2021 – May 2022 | | | | |
|---|---|---|---|---|
| Ranking | App | % market share | % urgent complaints | % aggressive debt collection complaints |
| 1 | Foxtrot | 21% ▲ | 8% | 12% |
| 2 | Delta | 4% ▲ | 34% ▲ | 43% |
| 3 | Juliet | 2% ▼ | 27% | 36% |
| 4 | Island | 2% ▲ | 10% ▲ | 33% ▲ |
| 5 | Kilo | 4% ▼ | 11% | 20% |
| 6 | Lima | 4% | 4% | 22% |
| 7 | Matcha | 2% | 33% | 15% |
| 8 | Numbus | 2% | 26% | 18% |
| 9 | Alpha | removed[2] | removed | removed |
| 10 | Beta | removed | removed | removed |
| 11 | Charlie | removed | removed | removed |
| 12 | Echo | removed | removed | removed |

8 most concerning apps = 41% of market

**Foxtrot**, although having grown its market share, **now has less concerning risk indicators**.

**Lima, Matcha and Numbus** were ranked lower and now have come onto the watchlist.

**Alpha, Charlie and Echo** were removed from Google Play Store but as of November 2022, they are once again on offer in Google Play Store.

# India pilot: How has the watchlist changed in a year?

| Jan 2020 – May 2021 | App |
|---|---|
| 1 | Alpha |
| 2 | Beta[1] |
| 3 | Charlie |
| 4 ▲ | Delta |
| 5 | Echo |
| 6 ▲ | Foxtrot |
| 7 | Golf |
| 8 | Hotel |
| 9 ▲ | Island |
| 10 ▲ | Juliet |
| 11 ▲ | Kilo |

| May 2021 – May 2022 | App |
|---|---|
| 1 ▲ | Foxtrot |
| 2 ▲ | Delta |
| 3 ▲ | Juliet |
| 4 ▲ | Island |
| 5 ▲ | Kilo |
| 6 | Lima |
| 7 | Matcha |
| 8 | Numbus |

[1] Beta is CashBean. In February 2022, the Reserve Bank of India canceled the registration certificate of CashBean's parent company. See: https://www.rbi.org.in/scripts/BS_PressReleaseDisplay.aspx?prid=53317.

## The market has slightly improved, but concerning signs remain:

The risk indicators of the top watchlist apps have improved. On average, urgent complaints have improved from 28% to 21% and aggressive debt collection complaints has improved from 36% to 25% - still high averages.

The market is now dominated by Foxtrot. It has improved its risk indicators, but they are not low enough yet (8% urgent complaints and 12% aggressive debt collection complaints).

Regarding the new apps on the watchlist: **Lima** is a new player with significant market share and high aggressive debt collection but low urgent complaints. **Matcha** and **Numbus** have higher risk indicators and sizable market shares.

## There are still concerning movements in the top apps in the market:

The NLP tool identified **Beta[1]** early on. It was removed from the Google Play Store, but it was only one of the higher-risk apps on the watchlist.

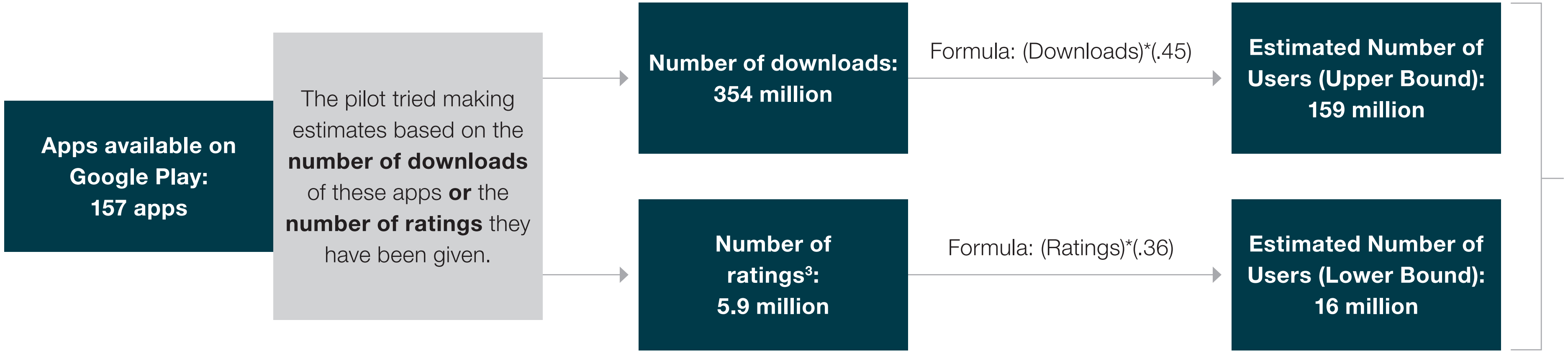
**Delta** grew from 3% to 4% in market share and its risk indicators remain high.

**Juliet** has a sizable market share and its risk indicators have worsened.

**Island** had poor risk indicators and was small but then grew in market share and still has high risk indicators.

# India pilot: Sizing the market

## How many users of digital lending apps are there?[1]




BUT this is a very wide range!

PLUS

The number of users may be under-estimated as many are unlikely to write Google Play reviews.

AND

The number of users may be over-estimated because some may take loans with multiple lenders.

So, it is uncertain whether the pilot likely over- or under-estimated the real number of users.

**Thus, market sizing is an important limitation of this tool.**

This estimate is based on May 2021 data.

[1] Globally, less than half (~45%) of downloaded (or installed) apps are used (Simform).

[2] Globally, only about 1/3 (36%) of using customers give feedback (CFIGroup).

[3] Note that our analysis searches Google Play reviews and Tweets for different *complaints*. This is different from a *rating* from 1 to 5 that a user might give an app on Google Play.


# IV. FURTHER RESOURCES

Photo by Vijay Pandey

# Relevant CGAP resources


Izaguirre, Juan Carlos, Denise Dias, Eric Duflos, Laura Newbury Brix, Olga Tomilova, and Myra Valenzuela, 2022. “Market Monitoring for Financial Consumer Protection.” CGAP toolkit.

Duflos, Eric, Mary Griffin, and Myra Valenzuela. 2021. ”Elevating the Collective Consumer Voice in Financial Regulation.” CGAP Working Paper.

Chalwe-Mulenga, Majorie, Eric Duflos, and Gerhard Coetzee. 2022. “The Evolution of the Nature and Scale of DFS Consumer Risks: A Review of Evidence.” Slide Deck. Washington, D.C.: CGAP.

Duflos, Eric, Jayshree Venkatesan, Amulya Neelam, and Sarah Stanley. 2021. “Digital Consumer Credit in India – Time to Take a Closer Look.” CGAP Blog post.

Duflos, Eric, Daryl Collins, Jayshree Venkatesan, and Juan Carlos Izaguirre. 2021. “Analyzing Social Media to Spot Digital Consumer Credit Risks in India.” CGAP Blog post.

Izaguirre, Juan Carlos. 2020. ”Making Consumer Protection Regulation More Customer-Centric.” CGAP Working Paper.

# TECHNICAL ANNEX.
# Steps to create NLP code for determined parameters, based on the India pilot


Photo by Temilade Adelaja

# Data scraping process

## Google Play

1. On Google Play Store we searched for the term "loan apps" and found 250 loan apps.

2. App links for these 250 apps were scraped using a scraping tool (Parsehub)

3. App IDs were extracted from the URLs scraped.
   Example –
   App URL: https://play.google.com/store/apps/details?id=com.kreditbee.android
   App ID: com.kreditbee.android

4. A Python code was used to crawl the Google Play Store and scrape the relevant data within a date range, using the "app IDs"

5. Reviews that were paid-for were stripped out.

## Twitter

1. From the app IDs extracted, we created hashtags using the following methods:
   App name (E.g. – Kreditbee)
   App name + "loan" (E.g. – Kreditbeeloan)
   App name + "app" (E.g. – Kreditbeeapp)

2. Few app names differed from their app IDs, hence we created hashtags from the app IDs as well, using the above method.
   Example:
   App name: "India AI Credit Cash loan app"
   App ID: "com.cashdrm.india"
   Hashtags - #cashdrm, #cashdrmloan, #cashdrmapp

Hence, a minimum of 3 hashtags were created for each app.

3. We didn't include the app names and app IDs whose names included common terms.
   Example: cashcredit, earlysalary, smallloan

4. All these hashtags were used in a Python code to scrape the Tweets.

5. By going through few tweets and referring to the previous time period issue-specific tweets, we found some more issue-specific tweets.

6. We used these new issue-specific hashtags to re-scrape the data.

With each iteration of scraping, we ended up with 21 new issue-specific hashtags.

Examples of issue-specific hashtags found – #digitallending, #fakeloanapps, #illegalloanapps, #fraudloanapps, #banloanapps

# Topic modeling and topic classification using natural language processing (NLP)

## Step 1: Topic Modelling

**Finding topics** by groups of keywords

We ran a code on the dataset in Python© to find which words related to complaints are likely to occur together.

The output is keywords, which helps us see emerging themes.

This step helps to find the key topics which, ultimately, the Topic Classification coding will use to detect trends on an ongoing basis.

*Note that it is important, but challenging, to analyze posts in different relevant local languages. See details of how we did this in Hindi at the end of this Annex.*

## Step 2: Topic Classification

**Refining topics**, informed by Step 1 results

Topic Classification works like a very complex word search on a Word document.

We then set "rules" for MeaningCloud© to find relevant entries.

The software scans the data set for entries that match the rules and classify them accordingly.

This step builds the software code which will be used on an ongoing basis to produce the results discussed below to monitor the market.

# NLP programming for consumer complaints

## Example: Inadequate Redress Mechanisms

**Example** '@credicxo Pls cancel my loan application otherwise give me loan . Its been 1 months submitting application. Cancel my loan application or desburse me loan . I return you loan in lockdown period. You promised me to give loan but you cheated me. Not replying to my mails/message/calls.

## Step 1: Topic Modelling

| Topic | Keywords | Theme |
|---|---|---|
| 0 | customer, service, call, response, team, send, mail, support, care, number | Recourse |

## Step 2: Topic Classification

| Category | Rule |
|---|---|
| Inadequate redressal mechanisms | call\|helpline\|callcenter\|callcentre\|mail\|email\|e-mail\|chat |
| > | AND "pick up" OR |
| Unresponsive complaints procedure | pickup\|pick\|centre\|center\|busy\|number\|disconnect\|same\|reply\|respond\|response\|answer\|automated\|robotic |

Pipeline (|) is same as "OR"

# NLP programming for consumer complaints (continued)

## Example: Fraud

**Example** I would rate even zero if i had a option... They will tale all our information and later on they says..!! It's out of service area..!! A fraud app dont ever go for it.!! If any thing happns with my personal details hopefull it would be a big mistake for you guys...!!

## Step 1: Topic Modelling

| Topic | Keywords | Theme |
|---|---|---|
| 3 | app, bad, give, fake, people, waste, download, fraud, star, install | Fraud |

## Step 2: Topic Classification

| Category | Rule |
|---|---|
| Fraud<br><br>><br><br>Fake app | fr[aou]+d\|fake\|fak\|scam\|install\|china\|chinese\|farji\|-farzi<br><br>AND<br><br>data\|information\|info\|details\|personal\|docu-ment\|samachar\|bechne\|bechke"customer\|cus-tomers\|person\|people\|log\|logon data\|inf[ro]*ma-tion\|info\|details" |

Pipeline (|) is same as "OR"

# NLP programming for consumer complaints (continued)

## Example: Lack of Transparency

**Example** That's true one of my friends took loan amount of 11000/- and total amount you charged 14500/- that only allowed one month, for that you charged 3500/- extra. He did make partial payment by due date and with extra charges. Where are you giving 3-6 months, you just giving one month frame and allowing partial pay and after you are charging extra amount for that as well. And who paid amount in more than 1 month you are stopping them from next loan or saying not serviceable area.

## Step 1: Topic Modelling

| Topic | Keywords | Theme |
|---|---|---|
| 3 | good, money, interest, fast, high, rate, loan, low, cash, recommend | Hidden Terms |

## Step 2: Topic Classification

<table>
<tr><th>Category</th><th>Rule</th></tr>
<tr><td>Lack of transparency</td><td>hide|extra|unexpected|unclear|unknown|high</td></tr>
<tr><td>></td><td>AND</td></tr>
<tr><td>Hidden Terms</td><td>fee|charge|penalty|intrerest</td></tr>
<tr><td></td><td>AND NOT</td></tr>
<tr><td></td><td>-hide</td></tr>
</table>

Pipeline (|) is same as "OR"

# Defining states of vulnerability

**Note:** It's important to note that this tagging of "vulnerability" is not static, such as a poverty level, but rather a state of vulnerability, such as losing a job or not being able to buy medication. Therefore, these should not be interpreted as "consumers who are vulnerable" but rather "consumers who report being in some state of vulnerability."

## States of Vulnerability

- I lost my job/I cannot find work/ I haven't been paid
- I am sick or need to take care of someone sick
- I cannot afford food or medicine
- Others

# Classification of reviews and tweets under states of vulnerability

| State of Vulnerability | Channel | Examples |
|---|---|---|
| Lost job/lack of work/not paid | Google Play | Worst Experience with collection team. Never take a loan from early salary. If your father died 3 months ago u have lost your job and your mother is hospitalised then also this guys do not have anything in mind just they say u too make the payment right now do whatever you want to do or our executive will visit you. This is the way they say to you. The worst experience ever |
| Can't afford food/medicine | Twitter | @cyber My issue is I took loan from a (CREDIME APPLICATION) amount 3000 available in Play Store but due to covid19 I lost my job,iam facing problem for daily food,support team created a group on my name as fraud ,they taken my contacts and Black mailing me please help sir |
| Direct/indirect sickness | Twitter | @IIFL_Finance @App_MyMoney<br><br>This is the 14-15th time I'm contacting you guys still no response from you. I want my refund as you took my EMI twice. I'm sick due to Covid and need money urgently for medical expenses.<br><br>Loan no SL2836928 |
| Others | Twitter | Waste App , waste of deposit money .Donno what they will gain by looting innocent people who install this app in tough situations in need of money instead of providing loan they are taking money and not giving any reply. We should all give complaint to cyber crime. |

# NLP software coding for vulnerable consumers

## Detecting states of vulnerability

| Topic (state of vulnerability) | Search rule[1] to find relevant entries | Example of relevant entry |
|---|---|---|
| Lost job/ lack of work/ not paid | (no\|lost\|koi\|nahi\|don't\|lock\|lockdown AND job\|work\|kaam\|salary\|income AND NOT working\|good\|nice\|people\|us\|illegal\|payday\|flex<br><br>\|ek) OR jobless // (no\|lost\|koi\|nahi\|don't\|lock\|lockdown AND job\|work\|kaam\|income\|salary) AND NOT working\|good\|nice\|illegal\|payday\|flex\|ek | "@AnilSinghvi_ @zeebusines @RBI #OperationHaftaVasooli @PayMeIndia<br><br>I have borrow small on payme India before lockdown. due to no work . no salary. I'm unable to pay my dues. But this agent are hack my all contacts and called everyone to harass<br><br>They spoiled my refutation" |

Topic Classification works like a very complex word search on a Word document

The software scans the data set for entries that match the rules and classify them accordingly

[1] The rule is simplified here for illustration purposes. It can be written with various parameters to refine search.

# Serial borrowers

## Data from January 2020 to May 2022

### Identifying share of serial borrowers:

| Channel | % of total reviews/tweets | % of positive sentiment[1] | % of total complaints | % vulnerable |
|---|---|---|---|---|
| Google Play reviews | 2% | 3% | 12% | 2% |
| Twitter | 0.2% | - | 5% | 1% |

[1] Positive sentiment only occurs in Google Play reviews.

### Identifying average loan sizes of serial borrowers (whenever a loan size can be detected):

| Channel[2] | % that are large loans | % that are medium loans | % that are small loans |
|---|---|---|---|
| Google Play reviews | 7.4% | 7.4% | 85.2% |

[2] Loan size can only be detected in Google Play reviews.

# NLP software coding for serial borrowers

## Detecting serial borrowers

| Topic | Search rule[1] to find relevant entries | Example of relevant entry |
|---|---|---|
| Serial borrowers | (Loan AND already\|before\|reason AND NOT received\|blog\|paydayloans\|post\|payday\|instant\|buddyloan\|gone\|bad\|update\|thousand\|illegal\|illigeal\|hack\|beware\|ilegal\|insurance\|hurt) OR (loan AND reject\|approve\|request\|declined\|cancel AND even\|declined\|application AND NOT received\|blog\|paydayloans\|post\|payday\|instant\|buddyloan\|gone\|bad\|update\|thousand\|illegal\|illigeal\|hack\|beware\|ilegal\|insurance\|hurt) | "I have already taken loan 2 times and mg both the time bills are paid on time. This 3rd time when I took loan again the loan is approved fast but the money hasn't come yet it is kept in showing disbursaing it's still not come yet such a bad service with a regular customer" |

[1] Note: rule is simplified here for illustration purposes. It can be written with various parameters to refine search.

# The complexities of topic modeling with Hindi script

There are Hindi script libraries but this is only part of the challenge of doing topic modeling in Hindi with Hindi script.  There are other things that need to be done to clean and prepare the data:

**Stemming** = Making sure that all the forms of a word are recognized.

Example:  am, are, is > be

In English, we use Porter's algorithm

We used snowballstemmer to do this in Hindi

**Tokenization** = Deciding what a word is.

Example: A sentence will be divided into single words

In English, we use NT LK/SpaCy

We used Indic NLP in Hindi

**Stop words** = Common terms to drop.

Example:  a, has, he, it.

In English, we use NLTK

We used database of stopwords manually created by multiple individuals to do this in Hindi

We then used Hindi LDA method to generate the topic modeling.

# Initial topic modeling results

## Twitter – Hindi language/Hindi script

Twitter = 73,485 tweets;

5,144 (7%) are Hindi language/Hindi script

| Key words from Hindi topic model | Translated | Number of tweets in which keywords are relevant | Topic (generalized) |
|---|---|---|---|
| | **Unambiguous customer complaint topics that are policy-relevant** | | |
| बैन ,एप्स ,पर्मानेन्टली ,सरकार ,जनता ,यहीं ,दुःखद ,करों ,बढ ,घटनाओं | Ban, Apps, Permanent, Government, Public, Here, Sad, Taxes, Increased, Events | 299 | Harassment (Moratorium by government) |
| कोर्ट ,सारे , धमकी , करके ,रिकवरी,कॉल ,बात ,गाली ,फ़ोन ,ले | Court, All, Threatening, By Doing, Recovery, Call, Talk, Abuse, Phone, Take | 226 | Harassment (Threatening calls) |
| कंपनियों ,ऑपरेशन ,बिज़िनेस ,ज़ी, स्टिंग ,भेजें ,ऑपरेशन... ,ध्यान , अनिल ,मुद्दे | Companies, Operations, Business, Zee, Sting, Send, Operation, Meditation, Anil, Issues | 42 | Fraud (Results of Sting operation on loan apps) |
| चाहिए , चुकी ,भेज ,1 ,धन्यवाद , कम्पनी ,मीडिया ,लिख ,देता ,जाता | Should, have, send, 1, thank you, company, media, write, give, go | 36 | Recourse (Customer service unresponsive) |
| जाने ,सबकी ,कांटेक्ट ,लिस्ट, इनकी ,बस ,मदद, वजह | Know, everyone, contact, list, their, bus, help, reason | 18 | Harassment (Contact list theft, threatening) |
| | **Ambiguous customer complaint topics that are policy-relevant** | | |
| करो ,लोग ,करे ,न्याय ,हम ,रही ,हफ्ता ,लोगों ,सुसाईड ,टॉर्चर | Do, people, do, do justice, we are, stay, week, people, suicide, torture | 1882 | Ambiguous (Harassment, customer complaints) |
| लोन ,नही ,कंपनी ,लोगो ,क्या , अब ,अगर ,लिया ,दे ,दिन | Loan, no, company, people, what, now, if taken, give, day | 1235 | Ambiguous (Inaccurate data, harassment) |
| वसूली ,वाला , दहला , फर्जी , जल्द, offc , हर , ट्वीट ,नोटिस | Recovery, Going, Dahala, Fake, Soon, offc, Her, Tweet, Notice | 42 | Ambiguous (Customer complaints) |
| केस ,आरबीआई ,महीने ,लूट ,फ्रॉड ,उड़ा ,आवाज ,ख़त्म ,क्रेडिट | Case, RBI, Month, loot, Fraud, Blow, Voice, Finish, Credit | 9 | Ambiguous (Ads, fraudulent apps) |

African Development Bank / Banque Africaine de Développement
African Development Fund / Fonds Africain de Développement

AFD

Bill & Melinda Gates foundation

British International Investment

Udenrigsministeriet
Ministry of Foreign Affairs of Denmark

European Investment Bank

Federal Ministry for Economic Cooperation and Development

giz

Global Affairs Canada

IDB | LAB

INVESTMENT FUND FOR DEVELOPING COUNTRIES

IFC International Finance Corporation

IFAD
Investing in rural people

AGENZIA ITALIANA PER LA COOPERAZIONE ALLO SVILUPPO

JOA
JERSEY OVERSEAS AID

KFW

KOICA Korea International Cooperation Agency

THE GOVERNMENT OF THE GRAND DUCHY OF LUXEMBOURG

Center for Inclusive Growth

mastercard foundation

MetLife Foundation

Ministry of Foreign Affairs and International Cooperation

Ministry of Foreign Affairs of the Netherlands

FMO
Entrepreneurial Development Bank

PROPARCO
GROUPE AFD

Sida

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Agency for Development and Cooperation SDC

Norad

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

UKaid
from the British people

UNCDF
Microfinance • Capital Development

UNDP

USAID
FROM THE AMERICAN PEOPLE

CGAP | cgap.org

@cgap

Facebook

LinkedIn