CASE STUDY

N.2 | OCTOBER 2024

FINANCIAL MARKET MONITORING VIA SOCIAL MEDIA AND WEB EXTRACTION ADVANCED ANALYTICS PLATFORM

SUPTECH WORKING PROTOTYPE DEVELOPED BY THE CAMBRIDGE SUPTECH LAB AND ITS PROJECT PARTNERS, THE SUPERINTENDENCE OF BANKING, INSURANCE AND PRIVATE PENSION FUNDS ADMINISTRATORS OF PERU AND FINANCIAL NETWORK ANALYTICS (FNA), WITH DATA DELIVERED BY WINNOW TECHNOLOGIES

This case study outlines the development of an artificial intelligence (AI) and machine learning (ML) -powered market monitoring prototype designed to enhance the supervisory capabilities of Peru's Superintendence of Banking, Insurance, and Private Pension Funds Administrators (SBS, for its acronym in Spanish). Building upon this initial prototype, with the support of Cambridge Suptech Lab, SBS and suptech vendor Financial Network Analytics (FNA) - with assistance from Winnow Technologies (Winnow)- developed a solution which, using Artificial Intelligence (AI) and Machine Learning (ML), allows for the segmentation and categorisation of information through advanced text analysis techniques applied to user posts on social media. These enhanced market monitoring tools extend SBS' capacity, surpassing basic sentiment analysis to discover trends, anomalies, and other significant patterns.

Cambridge Centre for Alternative Finance
UNIVERSITY OF CAMBRIDGE
Judge Business School

CAMBRIDGE SUPTECH LAB

www.cambridgesuptechlab.org

Project overview

The financial authorities responsible for market conduct supervision and financial consumer protection rely on intelligence gathered from various sources of information that capture users' experiences at every stage of their interaction with supervised institutions regarding financial products and services. This process encompasses operational data from users who engage with those products and services, reports issued by supervised institutions, qualitative and quantitative market research, complaints submitted by users to supervised entities, claims and complaints filed with the National Consumer Protection Authority, claims filed with SBS and user-generated content on social media, blogs, news sites, and other digital platforms.

The sheer volume of public data makes manual searching and analysis a time-consuming and complex task. To address this, SBS, within the framework of the project "Strengthening the Market Conduct Management Supervision Model", sought a suptech solution capable of automatically scraping, filtering and classifying a large volume of public, web-based data to identify potential market misconduct.

As part of the advanced text analysis tools used in this solution, topic modeling - a ML and AI technique- enhances financial supervision by uncovering underlying themes within large collection of text that may have been overlooked through manual review methods, making it easier to organise and interpret vast amounts of information and develop new taxonomies and relationships between keywords. Similarly, sentiment analysis examines opinions, sentiments, and emotions in text, classifying them as positive, negative, or neutral. These text analysis tools enable supervisors to focus on topics and patterns that consumers perceive as negative, potentially identifying emerging trends or shifts in market behaviour and providing a proactive tool to detect potential misconduct or issues before they escalate into major problems.

Through design sprints and agile iterations, the Lab collaborated with SBS and FNA to co-create a comprehensive solution that scrapes the data – with the help of subcontractor Winnow Technologies (Winnow).

The prototyped solution extracts detailed topics, matches relevant keywords, and assign sentiments. The results are presented in interactive dashboards, providing the authority with a quick overview of social media activity and enabling efficient drill-downs into specific topics or individual posts.

The working prototype leverages various AI and ML techniques, including natural language processing (NLP) methods for topic modeling, keyword tagging and sentiment analysis, using advanced models such as Bidirectional Encoder Representations from Transformers (BERT), Gensim and Generative Pre-trained Transformer (GPT). The output is displayed in a cloud-based graph database with filtering and visualisation features, such as interactive timelines, treemaps, bar charts, and tables. This solution demonstrates the value of AI and ML in the supervisory process, helping SBS shift towards more proactive and preventive market conduct oversight.

The project exemplifies the benefits of collaborative development between financial authorities, the Lab, and tech vendors, showcasing how joint efforts can swiftly produce advanced suptech solutions while achieving cost efficiency. According to the Suptech Generations framework, the solution enhances data collection through more refined web scraping and machine learning models, pushing the boundaries of supervisory technology.

Project partners

- **The Superintendence of Banking, Insurance and Private Pension Funds Administrators of Peru (SBS):** Financial authority that regulates and supervises financial institutions, insurance companies, and private pension funds administrators in Peru, ensuring market conduct supervision and financial stability.
- **Financial Network Analytics (FNA)**: Technology firm specialises in analytics and simulation, used by central banks, government authorities, and financial infrastructures globally.
- **Winnow Technologies** (subcontractor): Technology firm specialises in web-based data mining, natural language processing, sentiment analysis, topic modeling, and advanced analytics.

Challenges with the pre-existing tool

- **Restrictions to social media analysis:** Restrictions to social media analysis: The social media analysis efforts by SBS were limited in the depth of analysis, as the solution primarily relied on sentiment classification and basic topic categorisation, which was manually performed by a local vendor analyst. This restricted the system's ability to effectively classify, process, alert, and detect information related to potential market misconduct, reducing the overall efficiency of the analysis. Moreover, the lack of integration between social media and other sources, such as customer complaints, requires additional time and manual effort from supervisors to gather and analyse data, reducing overall efficiency. from supervisors.

- **Limited market options for specialised tools:** The tool was primarily designed for marketing analytics, focusing mainly on sentiment analysis and lacking specialised features needed for market conduct supervision. Since no tools tailored to these specific needs were available in the market, SBS faced challenges in adapting this tool to its specific needs.
- **Dependency on external providers:** SBS reliance on the vendor for configurations severely restricted its ability to customise the tool effectively. This dependency limited SBS's capacity to update classification methods, processing rules, alerts, or detection parameters independently. As a result, SBS struggled to adapt the tool to meet its evolving supervisory requirements efficiently.

Key features

- **Web scraping and data integration:** Automates data collection from various web and social media platforms, while also allowing the import of structured and unstructured data in various formats.
- **Centralised data warehouse:** Consolidates multiple data streams, providing a comprehensive view for analysis.
- **AI/ML based advanced text analysis:** Complementing AI with ML models that identify underlying themes and sentiment in web-based data through interpretation of relationships between words and their contextual meaning, helping to organise, understand and classify large information sets.
- **Advanced topic classification:** Uses a combination of supervised and unsupervised machine learning models for topic classification. The supervised models assign tags based on predefined keywords, while unsupervised models extract additional, more detailed topics based in the overall context of the collected information.
- **Interactive dashboards:** Visualises data and insights with the filtering and drill-down capabilities to support supervisory actions.

Benefits

- **Broad, automated social media monitoring and analysis:** The prototype developed automates web-based data collection and applies near real-time advanced analytics to provide better insights into financial institutions' market conduct and consumers' behaviors, preferences, and sentiment.
- **Prototype availability:** The prototype remains available for SBS to review the results of the analysis performed during its development.
- **Reliable, enhanced supervision:** AI/ML powered social media monitoring that enables in-depth text analysis, allowing supervisors not only to identify sentiments but also to detect emerging trends and potential misconduct, enhancing proactive market conduct supervision.

1. BACKGROUND AND SUPERVISORY CHALLENGES

SBS adopted the regulation and supervision of market conduct of financial institutions, insurance companies and private pension fund administrators as one of its core mandates. As part of this effort, SBS required supervised entities to maintain robust market conduct management as a fundamental aspect of their organisational culture and business strategy.

Supervising market conduct aims to ensure that firms implement sound business practices in their interactions with current and potential clients. This involves collecting extensive information from various sources, enabling supervisors to assess the consumer experience with the providers of financial services and products throughout the life cycle – from presale, to contracting, to execution, and to resolution.

In recent years, the explosion of public data sources has presented an opportunity to enhance financial sector supervision. Initially viewed supplementary, these public data sources were not analyzed in an aggregated manner from a market conduct perspective. However, as public data became more central to societal discourse, SBS recognised that leveraging this information could provide valuable insights to identify possible misconduct.

In 2021, SBS started to engage vendors to scrape and analyze social media data to identify potential issues, including inappropriate application of commissions, fees, and charges, as well as weaknesses in the design and operation of new and existing products and services. While this solution captured relevant information using algorithms and hashtags linked to market conduct issues, its analytical capabilities were mainly restricted to sentiment classification and a basic topic classification that needed to be analyze manually.

Moreover, the data generated by the vendor could not be automatically integrated with other information sources, such as customer complaints, requiring additional time and effort from supervisors to achieve a comprehensive market overview. The tool relied only on rule-based algorithms rather than advanced AI and ML models. As a result, data collection was prone to bias or incompleteness, making it difficult to accurately classify topics and detect previously unidentified misconduct patterns.

Consequently, supervisors expended significant time and effort addressing weaknesses in the data collection process, frequently coordinating with the vendor to adjust the tool's parameters and requesting supplementary reports. The lack of integration with internal complaints data further compounded these issues, limiting the ability to detect possible misconduct and other emerging market conduct risks.

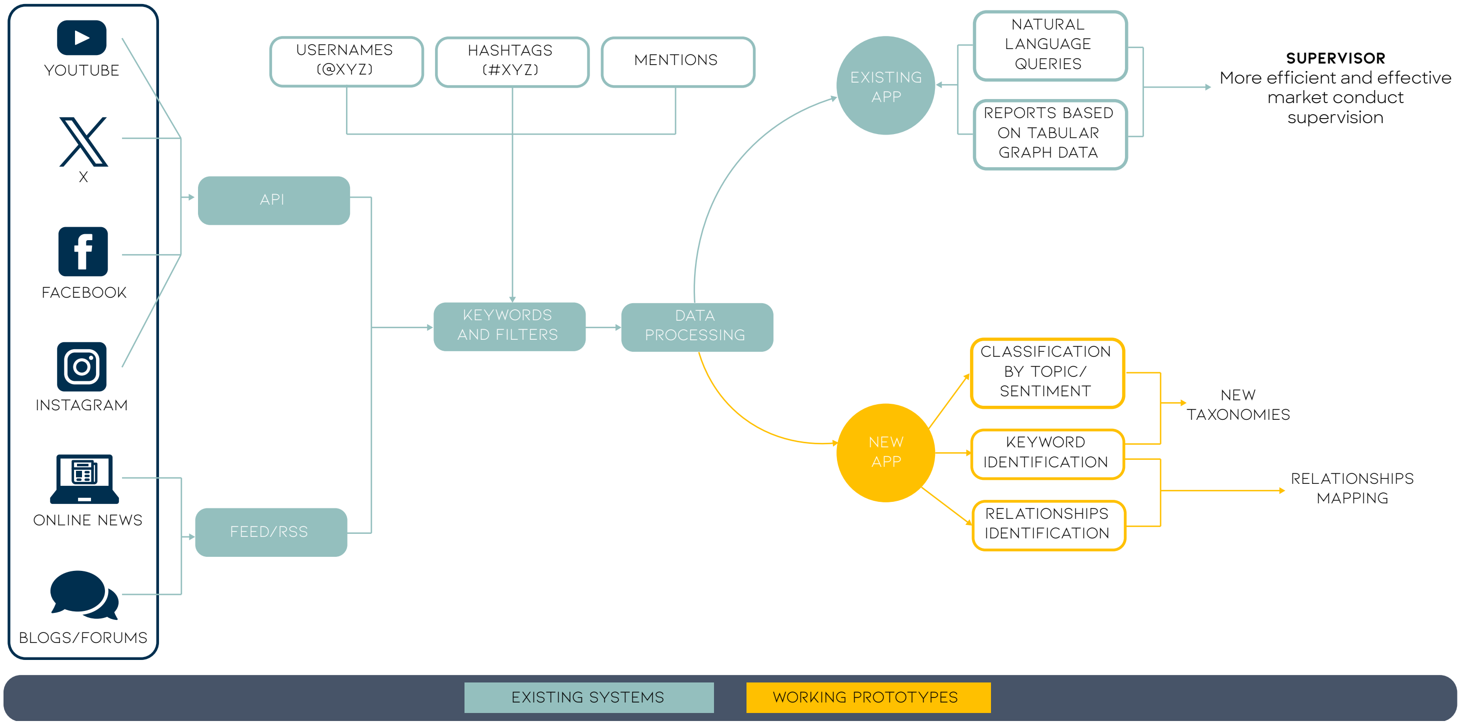
2. PROJECT CONCEPTUALISATION AND INCEPTION

In 2022, a team from SBS participated in the Lab's Capacity Building and Education (CB&E) online programmes, where they developed the proof of concept (POC) for a suptech application designed to enhance their supervisory capabilities in consumer protection.

SBS aimed to improve market conduct monitoring tools and processes by enhancing the analytics of the data scraped from public sources, utilising ML and AI to process that data, enhance analytical capabilities, and provide real-time market conduct alerts, allowing for more effective and dynamic supervision. This new technology would enable supervisors to automatically detect possible misconduct, set parameters, and identify key information more efficiently using natural language queries.

The Lab selected the SBS's POC for prototype development and supported the SBS team in refining their Project Charter and technical specifications. The goal of the prototype was to create a tool that would allow SBS to monitor social media for potential misconduct among Peru's five largest banks, enabling timely and efficient detection with minimal manual intervention. Figure 1 provides a schematic diagram of this solution, illustrating the existing web scraping components in teal and the advanced analytics capabilities of the working prototype in yellow.

FIGURE 1. SCHEMATIC DIAGRAM OF THE ENVISIONED FINANCIAL MARKET MONITORING PROTOTYPE



3. LEAN VENDOR SELECTION AND PROCUREMENT

In March 2023, the Lab procured the working prototype on behalf of the agency, executing a global competitive bidding via a Request for Proposal (RFP) and leading an expedited yet rigorous vendor selection process. An independent expert panel reviewed the anonymised bids from a global cohort of applicants, ultimately selecting the proposal by FNA.

The selection process began with the Lab and other University of Cambridge experts shortlisting responses to a request for expressions of interest (REOI) based on three criteria: (1) relevant experience (60%), technical and managerial expertise (30%), and adequate resourcing (10%). Firms that made the shortlist were subsequently issued an RFP. Proposals were reviewed by an independent panel of judges comprised of global experts and innovators. The evaluation criteria in this second phase emphasised topic responsiveness (65%), execution plan (25%), and innovative approach (10%).

Once FNA was selected as the vendor, the process underwent a no-objection review with SBS. The Lab and the University of Cambridge conducted through due diligence, formalising legal agreements related to data sharing and storage, intellectual property licensing, and public procurement. These terms were consolidated in a project agreement between the University and the vendor, which also included non-disclosure agreements (NDA).

By the end of April, following the completion of due diligence, the University of Cambridge contracted FNA. Once the vendor was onboarded, the Lab took charge of project management, overseeing the development and testing phases of the working prototype.

4. WORKING PROTOTYPE AGILE DEVELOPMENT

The project kicked off in June 2023, with the Lab facilitating the stakeholder and communication plan while leading design sprints with the cross-functional project team. FNA captured design documentation through user stories (Figure 2), technical requirements for desired features and functionality, and developed the working prototype according to the architecture diagram shown in Figure 3. During this phase, the project teams worked asynchronously, contributing comments, questions and suggestions through a collaborative, interactive and iterative process.

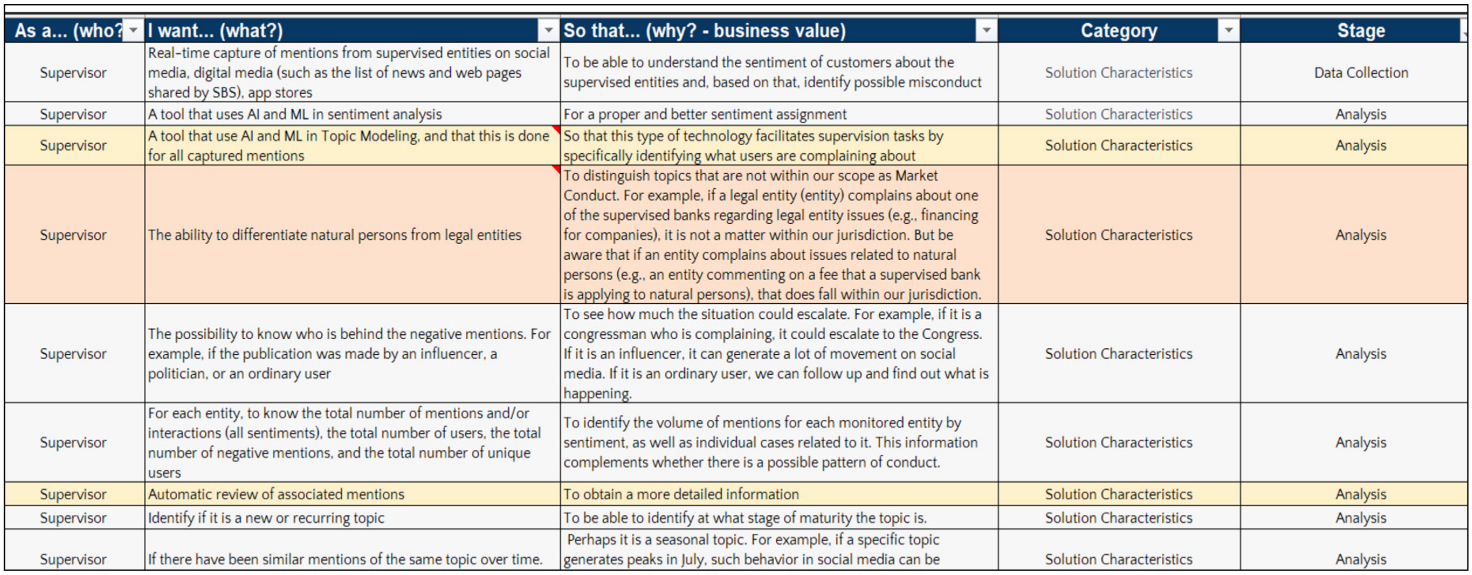
The working prototype for SBS focused on utilising AI and ML models to enable advanced text analysis by classifying social media posts through topic modelling, keyword identification and sentiment analysis. These tools informed the development of new taxonomies and relationship mapping. The aggregated data was then analysed, providing valuable insights into market conduct. The system enabled real-time monitoring and alerts, facilitating proactive supervision.

The Lab's prototype development process employs a unique approach to identifying potential challenges early in a project, ahead of a large procurement effort. This "front-loading" process helped SBS identify two critical issues early on.

First, it became apparent that simply layering analytics onto the pre-existing web scraping tool would not bypass the algorithms employed by social media sites to block data scraping. A more advanced web scraping tool was required. Second, extracting meaningful insights from social media proved to be challenging, as the connection between social media posts and market misconduct can be opaque. Negative social media posts often contain sparse information, foul language, or irrelevant content, such as emojis, images or symbols, making it difficult for both humans and machines to infer whether a complaint relates to actual market misconduct.

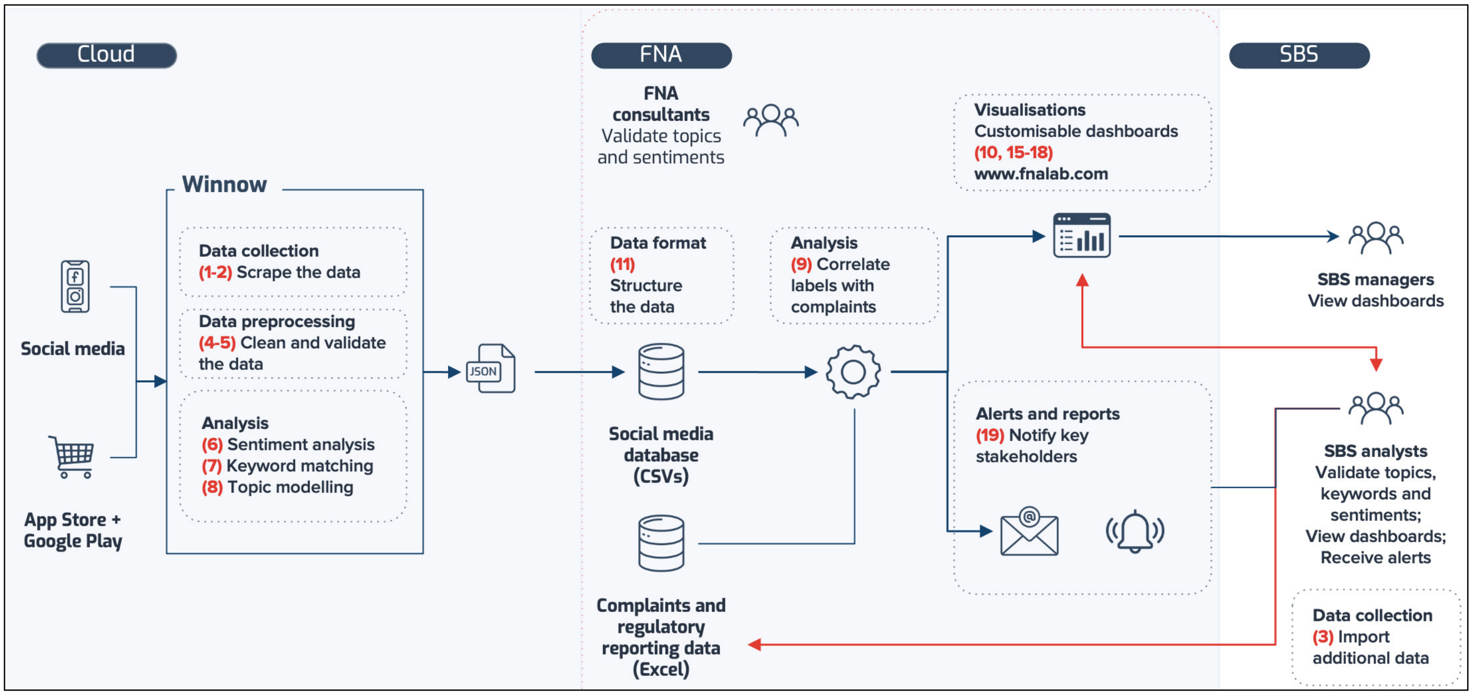
Any social media monitoring project of this kind should allocate ample time for iterative improvements to the NLP models. Iterative refinements of the models should continue beyond the initial development phase, as ongoing calibration will be necessary to adapt to changes in the environment.

FIGURE 2. USER STORIES



As a... (who?)	I want... (what?)	So that... (why? - business value)	Category	Stage
Supervisor	Real-time capture of mentions from supervised entities on social media, digital media (such as the list of news and web pages shared by SBS), app stores	To be able to understand the sentiment of customers about the supervised entities and, based on that, identify possible misconduct	Solution Characteristics	Data Collection
Supervisor	A tool that uses AI and ML in sentiment analysis	For a proper and better sentiment assignment	Solution Characteristics	Analysis
Supervisor	A tool that use AI and ML in Topic Modeling, and that this is done for all captured mentions	So that this type of technology facilitates supervision tasks by specifically identifying what users are complaining about	Solution Characteristics	Analysis
Supervisor	The ability to differentiate natural persons from legal entities	To distinguish topics that are not within our scope as Market Conduct. For example, if a legal entity (entity) complains about one of the supervised banks regarding legal entity issues (e.g., financing for companies), it is not a matter within our jurisdiction. But be aware that if an entity complains about issues related to natural persons (e.g., an entity commenting on a fee that a supervised bank is applying to natural persons), that does fall within our jurisdiction.	Solution Characteristics	Analysis
Supervisor	The possibility to know who is behind the negative mentions. For example, if the publication was made by an influencer, a politician, or an ordinary user	To see how much the situation could escalate. For example, if it is a congressman who is complaining, it could escalate to the Congress. If it is an influencer, it can generate a lot of movement on social media. If it is an ordinary user, we can follow up and find out what is happening.	Solution Characteristics	Analysis
Supervisor	For each entity, to know the total number of mentions and/or interactions (all sentiments), the total number of users, the total number of negative mentions, and the total number of unique users	To identify the volume of mentions for each monitored entity by sentiment, as well as individual cases related to it. This information complements whether there is a possible pattern of conduct.	Solution Characteristics	Analysis
Supervisor	Automatic review of associated mentions	To obtain a more detailed information	Solution Characteristics	Analysis
Supervisor	Identify if it is a new or recurring topic	To be able to identify at what stage of maturity the topic is.	Solution Characteristics	Analysis
Supervisor	If there have been similar mentions of the same topic over time.	Perhaps it is a seasonal topic. For example, if a specific topic generates peaks in July, such behavior in social media can be	Solution Characteristics	Analysis

FIGURE 3. SBS/FNA WORKING PROTOTYPE ARCHITECTURE DIAGRAM



5. THE APPLICATION

This project built upon SBS's foundational efforts to enhance its market conduct supervision capabilities, integrating advanced monitoring tools through social media and other channels. FNA partnered with Winnow Technologies, leveraging its robust capabilities in social media scraping which include extensive access to specific sites, and the ability to handle large volumes and diverse content.[1]

To effectively meet the objectives of the working prototype, FNA employed a combination of AI and ML methods that utilized both real-time and historical data. The comprehensive and ETL (Extract, Transform, Load) process was essential for transforming raw JSON data obtained from webscraping into a usable format. Following this, FNA applied various advanced analytical techniques:

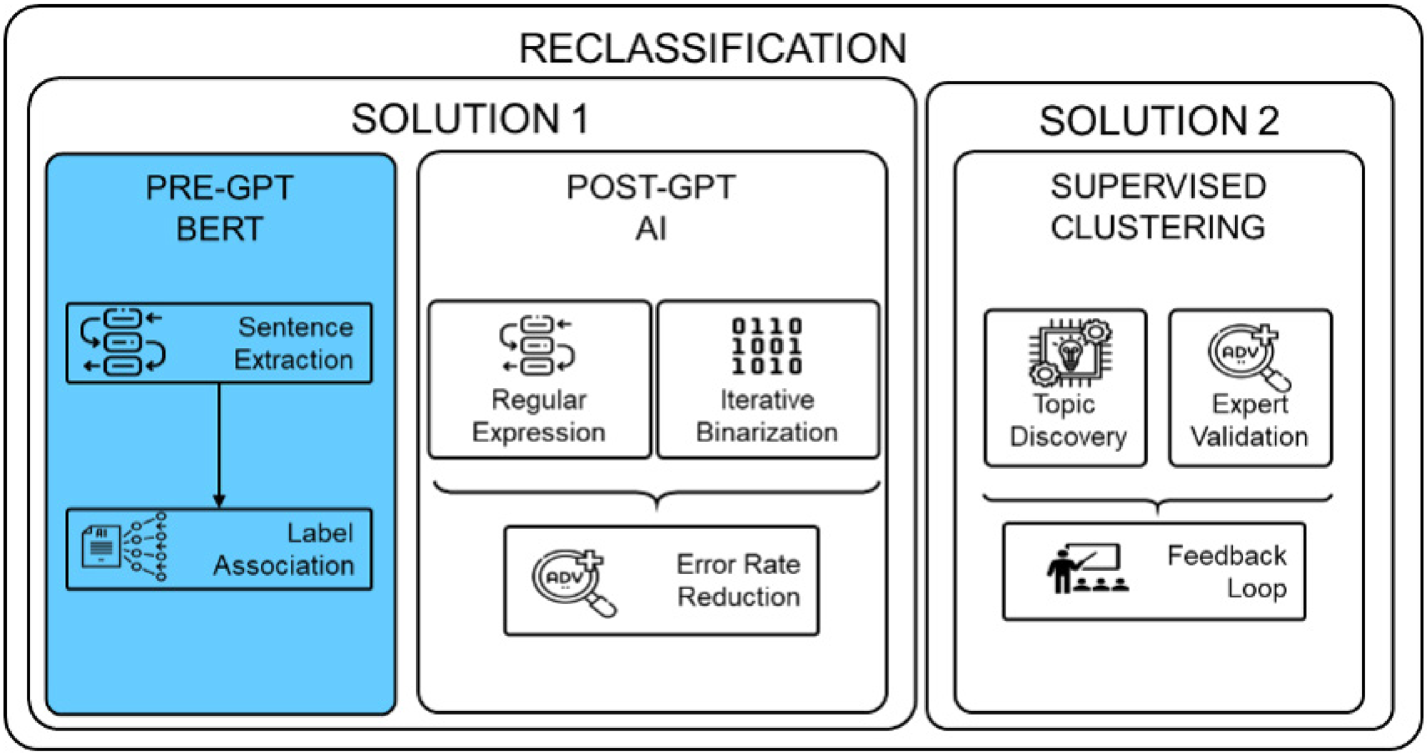
- **Sentiment analysis:** A multilingual, NLPtown/BERT-based model fine-tuned on about 50,000 social media post in Spanish to assign sentiments (Positive, Negative, or Neutral) to each one based on its contents. This model is supervised, meaning it was trained on previously labeled data to ensure accuracy and relevance in its sentiment assessments.
- **Labeling and categorization:** SBS supervisor provided sets of labels and categories they want to use for grouping the content. The Natural Language Toolkit (NLTK) was utilized for text preprocessing, including normalization and stemming, to enhance machine readability. Regular Expressions (RegEx) were also employed for text manipulation. Utilizing NLP models, both topic assignment and topic generation were performed to align content with an expanded list of topic-specific labels. This approach facilitates accurate classification, enabling efficient organization and interpretation of text data.
- **Detailed label generation:** This component was more complex, involving multiple steps to ensure precision:
  - **Unsupervised learning:** To support effective label generation, unsupervised topic modeling techniques, such as Latent Dirichlet Allocation (LDA), are utilized. This step uncovers underlying themes in data, facilitating the subsequent processes.
  - **Topic Generation:** BERT (Bidirectional Encoder Representations from Transformers) was employed for topic generation due to its ability to understand context and semantics, allowing for the extraction of meaningful themes from the corpus.
  - **Manual review:** A human-in-the-loop process was implemented for selecting the topics of interest and sampling data for manual label assignment By ensuring alignment with the specialized expertise of the reviewers. This maintains domain relevance and enhances the overall quality of the training sets.
  - **Supervised modeling:** A supervised model then processed the full population of posts using the refined labels, providing a structured classification of the detailed topics.

[1] See Cambridge SupTech Lab (2024b), Next-Generation AI-Powered, Chatbot-Supported Complaints Management System, Case Study N.3, forthcoming, Cambridge: Cambridge Centre for Alternative Finance (CCAF), University of Cambridge.

- **Content summarisation:** The BERT Extractive Summarizer extracted key sentences from the posts related to each topic, followed by GPT generating concise summary bullet points (labels). These summaries provided clear and actionable insights into the nature of discussions within each topic.
- **Visualisation:** Results from these processes were visualized through a cloud-based graph database that included interactive timelines, tree maps, bar charts, and tables, enabling dynamic and intuitive access to the data.
- **Classification**: For each identified topic, FNA assigned relevant social media posts and calculated both the number and percentage of posts per topic. This structured approach not only clarified the scale and specifics of each topic but also enhanced the monitoring and tracking of market conduct-related discussions across platforms.

This layered and methodical approach to data analysis ensured a comprehensive understanding of market conduct issues, providing SBS with the tools needed for proactive and informed supervision.

FIGURE 4.. SBS/FNA RECLASSIFICATION SOLUTIONS



Two potential solutions were explored to reclassify posts using a more granular set of topic labels. Both methods would have required manual validation and iteration to improve the results.

Solution 1: Match bullet points back to posts

The first solution focused on maintaining the summary bullet points while matching them back to the original posts. Two approaches could have been considered:

A. Match relevant sentences (Before GPT)

Using BERT Summarizer, this approach would combine and condense the content of all posts within a specific topic by removing irrelevant and repeated material, without generating new text. For each relevant sentence, exact matches within the content of one or multiple original posts would be identified. However, this process would not work in reverse; posts containing only irrelevant or repeated content might not be mapped back to relevant sentences.

The tool could have assigned an index to each relevant sentence before applying GPT. When using GPT to generate summary bullet points, the tool would also modify the prompt to ask GPT to provide the indices of the relevant sentence(s) used for each bullet point.

This method would ensure the ability to track exactly which original posts and relevant sentences led to the summary bullet points, producing accurate matches. However, there was a risk of missing some original posts, and it was uncertain whether GPT could handle this type of modified prompt without issues.

B. Match summary bullet points (After GPT)

Alternatively, the output generated by GPT could have been matched back to the original posts using a combination of similarity matching and regular expression. This approach could have been implemented and tested more quickly. However, because GPT reformulates and shortens relevant sentences, matching bullet points to original posts would have resulted in higher error rates. As a result, several iterations would have been needed to lower the error rate to an acceptable level, potentially taking longer than Solution **1A**.

In both Solution **1A** and **1B**, there would have been limited control over which specific bullet points GPT generated based on the relevant sentences, requiring manual intervention to implement feedback.

In both **Solution 1A** and **1B**, code cannot control which specific bullet points GPT generates based on the relevant sentences, which limits the ability to implement any feedback without manual intervention.

Solution 2: Create new granular subtopics

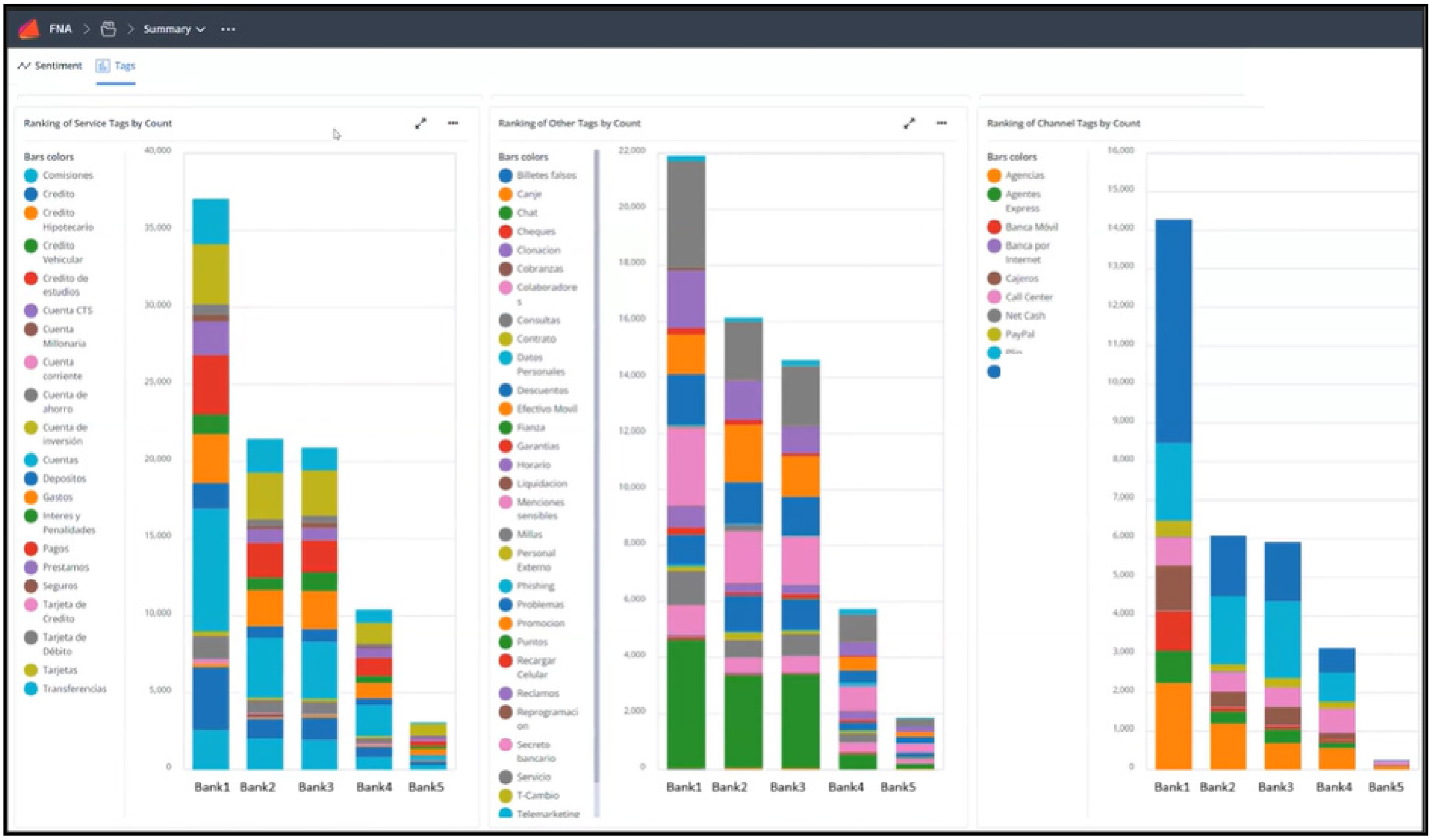
SBS ultimately selected this option as a more stable, long-term solution. This approach eliminated the BERT Summarizer and GPT steps entirely, instead focusing on creating new, more granular subtopics in place of summary bullet points. The method used for this solution was similar to the one employed to create the initial topics: a mix of unsupervised learning to identify clusters, manual review of a sample of posts to assign relevant subtopics, and supervised learning to assign subtopics to the remaining population of posts.

A subject matter expert from SBS collaborated with FNA to validate the subtopics and provided well-defined labels for a large sample of posts, which served as the training dataset. This approach gave SBS more flexibility in defining the subtopics and incorporated validation directly into the workflow.
While this solution required more manual work at the beginning, as well as occasional reviews, it offered SBS greater stability and control. It ensured that the subtopics aligned with ongoing discussions on social media and app stores, allowing for better adaptation to evolving market conditions.

The Summary dashboard displays the number of posts as they relate to keywords provided by SBS. The graph on the left displays posts relating to products such as credit, deposits, and loans. In the middle, the graph displays posts that relate to relevant tags provided by SBS relating to collections, promotions, personal data, and other data that may provide additional insights into consumer sentiments. The graph on the right displays the posts by the channels defined by SBS for monitoring and analysis, including call centers, branches or payment platforms.

The Sentiment dashboard displays data scrapped from social media for five large

FIGURE 5. FNA´S SUMMARY DASHBOARD



financial institutions during the first 6 months of 2023. Nearly 50,000 posts were analysed for Bank1, showing predominantly negative sentiments. The graph shows that a notable spike in negative posts took place in March, which informs the supervisory areas that warrant further investigation.

The Monthly Monitoring dashboard provides a closer look at each individual bank, with an overview of a bank's activity over a specified period, in this case 6 months. At the top of the dashboard, there is a notable spike in activity during March, consistent with Figure 4.

FIGURE 6. SENTIMENT ANALYSIS DASHBOARD

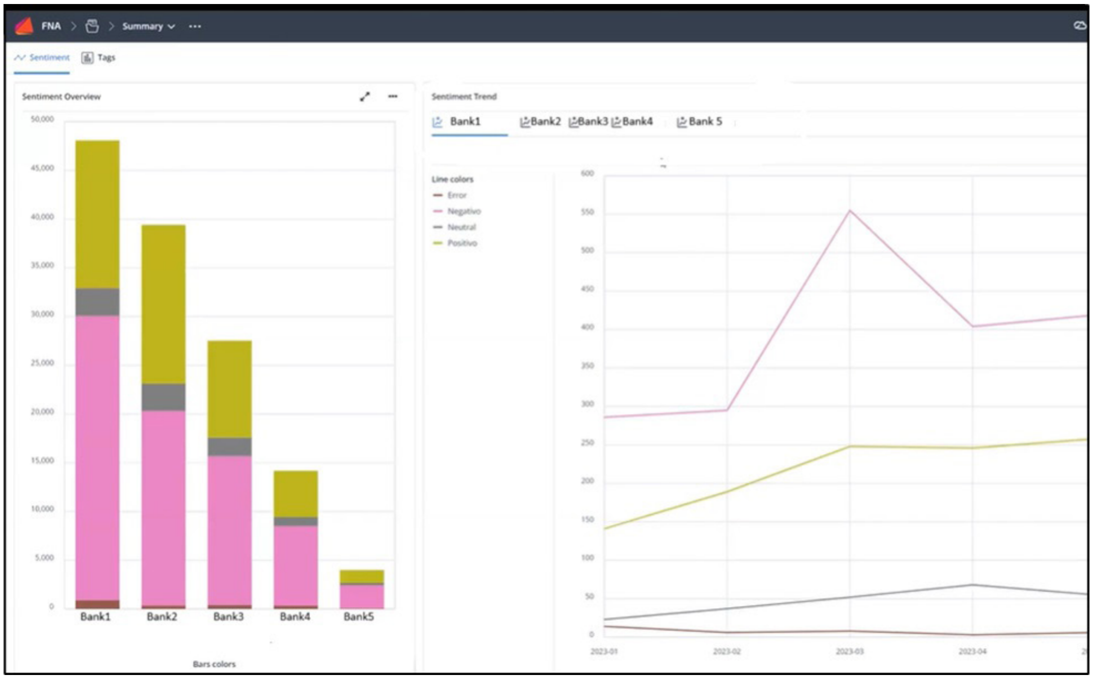
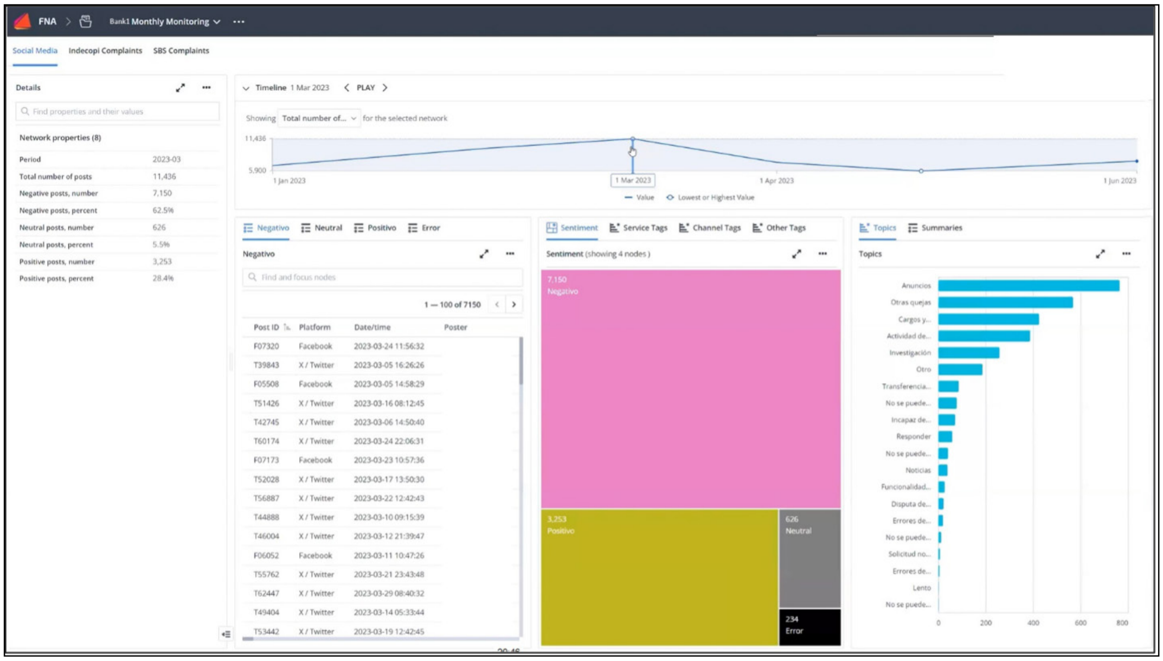


FIGURE 7. FNA'S MONTHLY MONITORING DASHBOARD FOR INDIVIDUAL BANKS



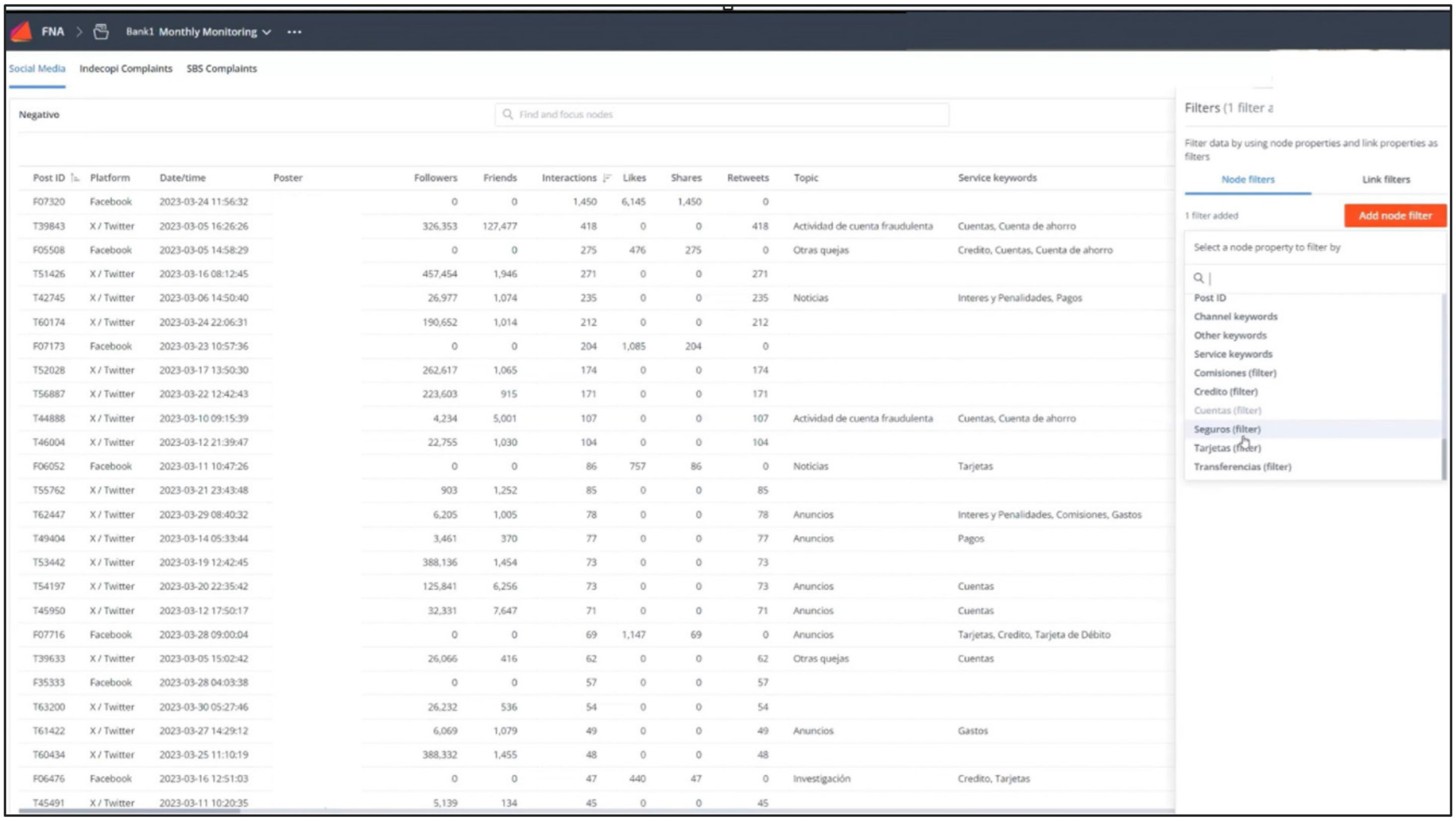
Network properties (8)

Period	2023-03
Total number of posts	11,436
Negative posts, number	7,150
Negative posts, percent	62.5%
Neutral posts, number	626
Neutral posts, percent	5.5%
Positive posts, number	3,253
Positive posts, percent	28.4%

Negativo

Post ID	Platform	Date/time	Poster
F07320	Facebook	2023-03-24 11:56:32	
T39843	X / Twitter	2023-03-05 16:26:26	
F05508	Facebook	2023-03-05 14:58:29	
T51426	X / Twitter	2023-03-16 08:12:45	
T42745	X / Twitter	2023-03-06 14:50:40	
T60174	X / Twitter	2023-03-24 22:06:31	
F07173	Facebook	2023-03-23 10:57:36	
T52028	X / Twitter	2023-03-17 13:50:30	
T56887	X / Twitter	2023-03-22 12:42:43	
T44888	X / Twitter	2023-03-10 09:15:39	
T46004	X / Twitter	2023-03-12 21:39:47	
F06052	Facebook	2023-03-11 10:47:26	
T55762	X / Twitter	2023-03-21 23:43:48	
T62447	X / Twitter	2023-03-29 08:40:32	
T49404	X / Twitter	2023-03-14 05:33:44	
T53442	X / Twitter	2023-03-19 12:42:45	

FIGURE 8. FNA'S FILTERABLE POST DASHBOARD



Negativo

Post ID	Platform	Date/time	Poster	Followers	Friends	Interactions	Likes	Shares	Retweets	Topic	Service keywords
F07320	Facebook	2023-03-24 11:56:32		0	0	1,450	6,145	1,450	0		
T39843	X / Twitter	2023-03-05 16:26:26		326,353	127,477	418	0	0	418	Actividad de cuenta fraudulenta	Cuentas, Cuenta de ahorro
F05508	Facebook	2023-03-05 14:58:29		0	0	275	476	275	0	Otras quejas	Credito, Cuentas, Cuenta de ahorro
T51426	X / Twitter	2023-03-16 08:12:45		457,454	1,946	271	0	0	271		
T42745	X / Twitter	2023-03-06 14:50:40		26,977	1,074	235	0	0	235	Noticias	Interes y Penalidades, Pagos
T60174	X / Twitter	2023-03-24 22:06:31		190,652	1,014	212	0	0	212		
F07173	Facebook	2023-03-23 10:57:36		0	0	204	1,085	204	0		
T52028	X / Twitter	2023-03-17 13:50:30		262,617	1,065	174	0	0	174		
T56887	X / Twitter	2023-03-22 12:42:43		223,603	915	171	0	0	171		
T44888	X / Twitter	2023-03-10 09:15:39		4,234	5,001	107	0	0	107	Actividad de cuenta fraudulenta	Cuentas, Cuenta de ahorro
T46004	X / Twitter	2023-03-12 21:39:47		22,755	1,030	104	0	0	104		
F06052	Facebook	2023-03-11 10:47:26		0	0	86	757	86	0	Noticias	Tarjetas
T55762	X / Twitter	2023-03-21 23:43:48		903	1,252	85	0	0	85		
T62447	X / Twitter	2023-03-29 08:40:32		6,205	1,005	78	0	0	78	Anuncios	Interes y Penalidades, Comisiones, Gastos
T49404	X / Twitter	2023-03-14 05:33:44		3,461	370	77	0	0	77	Anuncios	Pagos
T53442	X / Twitter	2023-03-19 12:42:45		388,136	1,454	73	0	0	73		
T54197	X / Twitter	2023-03-20 22:35:42		125,841	6,256	73	0	0	73	Anuncios	Cuentas
T45950	X / Twitter	2023-03-12 17:50:17		32,331	7,647	71	0	0	71	Anuncios	Cuentas
F07716	Facebook	2023-03-28 09:00:04		0	0	69	1,147	69	0	Anuncios	Tarjetas, Credito, Tarjeta de Débito
T39633	X / Twitter	2023-03-05 15:02:42		26,066	416	62	0	0	62	Otras quejas	Cuentas
F35333	Facebook	2023-03-28 04:03:38		0	0	57	0	0	57		
T63200	X / Twitter	2023-03-30 05:27:46		26,232	536	54	0	0	54		
T61422	X / Twitter	2023-03-27 14:29:12		6,069	1,079	49	0	0	49	Anuncios	Gastos
T60434	X / Twitter	2023-03-25 11:10:19		388,332	1,455	48	0	0	48		
F06476	Facebook	2023-03-16 12:51:03		0	0	47	440	47	0	Investigación	Credito, Tarjetas
T45491	X / Twitter	2023-03-11 10:20:35		5,139	134	45	0	0	45		

A closer look at the specific posts that make up the Monthly Monitor graph reveals many posts are advertisements, however notable posts by users can also show discussions relating to charges, fees, or potential fraudulent account activity.

This dashboard can be filtered by specific keywords, and be sorted by the level of interactions, such as the total number of reposts, likes, and shares. In this example, a news agency's post ranks highest on X (formerly Twitter). The second most interacted-with post comes from an individual user.

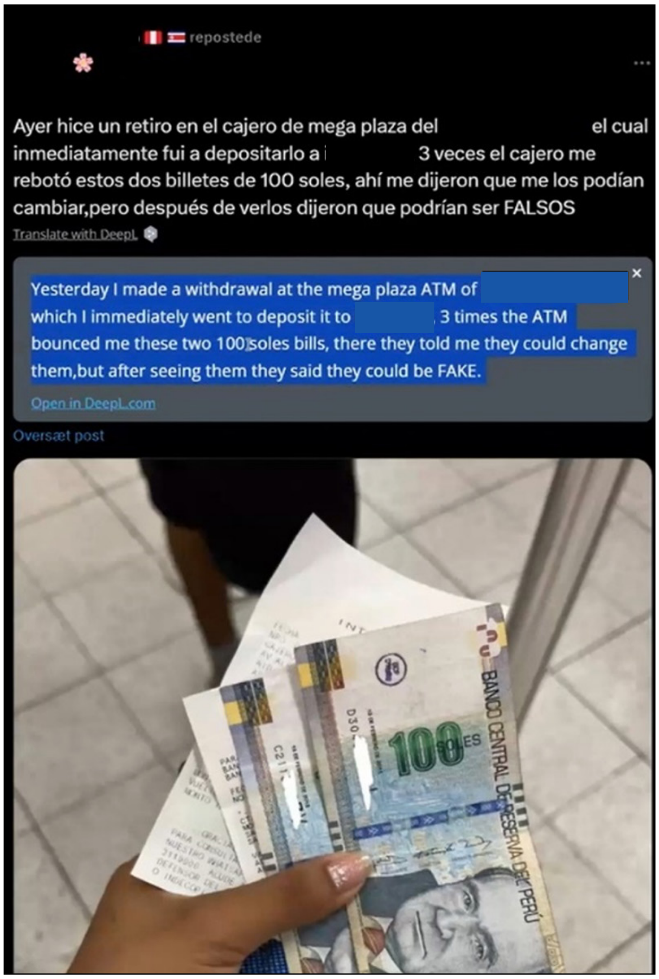
Users can drill down further to view the contents of each post. This following example shows an ATM withdrawal issue with concerns over potentially fake bills dispensed by the supervised entity.

The SBS Market Conduct department supervises 64 entities. The working prototype focused on scraping and analyzing data from five banks. A total of 341,838 comments were collected from seven platforms, including Facebook, X (formerly Twitter), Instagram, the Apple App Store, and Google Play Store.

FNA developed three major and eight minor prototype versions through iterative quality enhancements. Each new version refined the data filters, improved the topic modeling, and enhanced the user interface, building on the capabilities of the previous versions. The final result included 53 widgets across 20 pages on 8 interactive dashboards, enabling users to efficiently analyze hundreds of thousands of social media posts.

These dashboards allowed SBS to visualize and extract insights from the vast data sets collected, providing a more comprehensive view of market conduct trends. This advanced tool not only enhanced SBS's ability to monitor market conduct but also helped identify potential issues in real-time, improving the overall supervisory process.

FIGURE 9. SAMPLE POST



repostede

Ayer hice un retiro en el cajero de mega plaza del [redacted] el cual inmediatamente fui a depositarlo a i[redacted] 3 veces el cajero me rebotó estos dos billetes de 100 soles, ahí me dijeron que me los podían cambiar,pero después de verlos dijeron que podrían ser FALSOS

Translate with DeepL

Yesterday I made a withdrawal at the mega plaza ATM of [redacted] which I immediately went to deposit it to [redacted] 3 times the ATM bounced me these two 100 soles bills, there they told me they could change them,but after seeing them they said they could be FAKE.

Open in DeepL.com

Oversæt post

TABLE 1. FNA NETWORKING PROTOTYPE RESOURCES

WEB SCRAPING BY WINNOW	Platforms scraped	App Store Google Play Facebook Instagram X (Twitter)
	Comments scraped	341,838
ANALYSIS BY FNA	Sentiment analysis	NLPtown/BERT-base-multilingual-uncased-sentiment
	Key word tagging	Natural Language toolkit (NLTK) Regular Expression (RegEx)
	Topic modeling	Gensim Generative Pretrained Transformer (GPT) Latent Dirichlet Allocation (LDA) BERT Extractive Summarizer Classification machine learning models
VISUALISATION BY FNA	Dashboards	53 Widgets 20 pages 8 Dashboards

6. THE IMPACT

Having an AI/ML tool to monitor social media and other public sources, such as app stores, blogs, and news sites, allows for more proactive identification of trending topics concerning consumers of supervised entities. It also facilitates analysis from the perspective of market conduct supervision, while generating reports or alerts that can be used to prioritise and strengthen investigations or on-site inspections by leveraging:

- Automated data collection and reporting
- Near real-time capture and analysis of large data volumes
- Efficient keyword and hashtag searches to target specific issues
- Automatic integration of data analysis for improved information management
- Customised dashboards and reports for tailored insights
- Integration of information from multiple sources, offering a broader perspective

With this technology, financial authorities will be able to:

- **Strengthen risk-based supervision** by using automated monitoring through advanced text analysis tools to report and categorize real-time information about market conditions and emerging consumer risks. This allows agencies to make more informed decisions and manage risk effectively.
- **Enhance proactive and preventive action** through proactive monitoring achieved by advanced topic classification ensures that potential risks are identified, analyzed, and addressed before they escalate.
- **Receive near real-time updates** on processed unstructured data regarding specific products, issues, or entities, helping supervisors predict potential misconduct and intervene when risks become imminent (e.g., through on-site examinations).
- **Maintain a continuous influx of data**, providing early warning signals of potential financial misbehavior or reputational harm.

> “ *The availability of timely and comprehensive textual data enables us to stay ahead of emerging trends and potential issues within the consumer landscape. By harnessing this data, we can identify patterns, anomalies, and areas of concern in real-time, enabling us to take proactive supervisory measures. Additionally, applying advanced natural language processing techniques—such as topic modeling—allows us to condense large volumes of textual data into meaningful, actionable insights. This empowers us to prioritise and focus our supervisory efforts on areas with the most significant impact.*
>
> *Sentiment analysis provides another valuable layer of understanding by gauging public sentiment towards various products or entities. This analysis helps us identify potential misconduct and assess the effectiveness of current supervisory measures based on public perception.*
> *By leveraging these tools and techniques, we will significantly enhance the precision and effectiveness of our supervisory actions. These data-driven approaches ensure that our efforts are evidence-based, responsive to the evolving landscape, and ultimately focused on preserving consumer interests more effectively.* ”
>
> *Nicolas Tirado Vilela*
> *Market Conduct Analyst at SBS*

7. WHAT'S NEXT

After the prototype was tested, SBS confirmed that the solution accelerates their digital transformation by clearly demonstrating the value of integrating artificial intelligence and machine learning into the supervisory process. The prototype exercise also helped them better understand the necessary steps for moving into production. The Lab's approach is specifically designed to avoid vendor lock-in, ensuring that SBS retains flexibility.

As part of the Lab's Application Incubation program, agencies like SBS are provided with several options following the delivery of a prototype. They can:

1. Continue working with the vendors that developed the prototype.
2. Contract different vendors to further refine or develop the solution.
3. Deploy in-house resources to advance and maintain the solution independently.

SBS is currently evaluating its specific pathway to production, carefully considering which approach will best suit their operational needs and long-term strategic goals.

PROJECT PARTNERS

Superintendencia de Banca, Seguros y AFP (SBS) of Peru

SBS is the agency in charge of regulating and supervising the financial institutions, insurance companies and private pension fund administrators in Peru, as well as preventing and detecting money laundering and terrorism financing. Its main objective is to safeguard the interests of consumers and users of the mentioned companies, ensure their proper functioning, preserve financial stability and integrity, and ensure adequate market conduct.

FNA

FNA is a leader in advanced network analytics and simulation. Its software is used to uncover hidden connections and anomalies in large, complex datasets, to predict the impact of stress events, and to optimally configure financial systems and infrastructure. FNA is trusted by the world's largest central banks, government authorities, commercial banks and financial infrastructures.

Winnow Technologies Inc.

Winnow Technologies (Winnow) is a vendor that specialises in web-based data mining tooling, natural language processing and advanced analytics to assist public agencies in fulfilling their mandates to citizens and support the development of inclusive, sustainable and resilient markets, economies, and societies. The tools developed and deployed by Winnow allow the oversight of regulated firms and unregulated activities by scanning the web, social media, company reports and other communications to flag potential violation of policy and regulations, conduct sentiment analysis, and correlate collected information for supervisors on an ongoing basis.

About the Cambridge SupTech Lab

The Cambridge SupTech Lab accelerates the digital transformation of financial supervision to nurture resilient, transparent, accountable, sustainable, and inclusive financial sectors.

The Lab catalyses the scalable integration of innovative technologies, data science and agile methodologies by supervisory authorities to address the enduring and emerging challenges of the rapidly changing financial landscape. Through the Lab, financial authorities have championed the adoption of advanced suptech solutions that tackle critical issues such as financial crime, fraud, exclusion, climate change enablers, consumer protection, and artificial intelligence biases.

The Lab is hosted at the Cambridge Centre for Alternative Finance (CCAF) at the Cambridge Judge Business School, and leverages foundational intellectual property and know-how from the RegTech for Regulators Accelerator (R²A).

DISCLAIMER

The mention of specific companies, manufacturers, or software does not imply that they are endorsed or recommended by the Cambridge SupTech Lab in preference to others of a similar nature that are not mentioned.

SUGGESTED CITATION

Cambridge SupTech Lab (2024), Financial Consumer Protection Suite with Web Scraping and Machine Learning Based Analysis, Case study n.2, Cambridge: Cambridge Centre for Alternative Finance (CCAF), University of Cambridge. Available at www.cambridgesuptechlab.org

AUTHORS

Simone di Castri, Matt Grasser, and Nathalie Lenehan

DESIGN

Dayna Donovan

ADDITIONAL CONTRIBUTORS

Kalliope Letsiou, Susu Smaili