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/machine learning (AI/ML)-powered market monitoring prototype designed to enhance the supervisory capabilities of the Peruvian Superintendencia de Banca, Seguros y AFP (SBS). The solution designed and developed by the Cambridge SupTech Lab with suptech vendor Financial Network Analytics (FNA) and the SBS, integrates social media data scraped by Winnow Technologies (Winnow), sentiment analysis, topic modeling, segmentation, categorisation and other advanced ML techniques to discover trends, anomalies, and other significant patterns for market conduct supervision.

Cambridge
Centre
for Alternative
Finance
UNIVERSITY OF
CAMBRIDGE
Judge Business School

CAMBRIDGE **SUPTECH LAB**

www.cambridgesuptechlab.org

Project overview

The Superintendencia de Banca, Seguros y AFP (SBS) is the authority responsible for financial consumer protection market conduct supervision in Peru. To gain a comprehensive understanding of financial services users' experiences and their interactions with supervised institutions, the SBS relies on intelligence from a variety of sources. These include complaints data from users, periodic compliance reports issued by supervised institutions, qualitative and quantitative market research, complaints submitted to supervised entities, and claims filed with both the national consumer protection authority and the SBS itself. Recently, user-generated content on social media, blogs, news sites, and other digital platforms has also become a valuable source of insights.

Given the vast and growing volume of public data generated daily, manually searching, analysing, and extracting actionable insights is highly resource-intensive and complex. To enhance their off-site supervisory and market surveillance capabilities, the SBS has utilised a suptech solution that could automatically scrape, filter, and classify large volumes of web-based public data. This solution would sift out irrelevant content and categorise posts in real-time, enabling the efficient identification of potential market misconduct signals.

The SBS sought the support of the Lab to enhance the detection of emerging risks, anomalies and other patterns through advanced analytics of the scraped data, allowing a more swift and accurate response to market conduct issues.

The prototype developed by Financial Network Analytics (FNA) – with support from Winnow Technologies (Winnow), a suptech vendor that specialises in web and social media scraping, topic modeling and sentiment analysis[1] – leverages advanced deep learning models including Bidirectional

Encoder Representations from Transformers (BERT) to examine opinions, sentiments, and emotions in text, classifying them as positive, negative, or neutral. Gensim, a Python library designed for topic modeling, and Generative Pre-trained Transformer (GPT) pre-trained on large volumes of text in an unsupervised manner, then fine-tuned on specific tasks to label data.

These combined techniques enhance financial supervision by uncovering underlying themes within large collections of text that may have been overlooked through manual review methods. Vast amounts of information are easily organised and interpreted, and new taxonomies and relationships between keywords are defined. These 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 approach to detect potential misconduct or issues before they escalate into major problems.

The output is displayed in a cloud-based graph database with filtering and visualisation features, such as interactive timelines, tree maps, bar charts, and tables. Tree maps are a data visualisation technique used to display hierarchical data using nested rectangles, representing large amounts of information in a compact form. These are particularly useful for showing proportions and relationships within the data. This solution demonstrates the value of AI/ML in the supervisory process, helping SBS shift toward more proactive and preventive market conduct oversight.

The project exemplifies the benefits of collaborative development between financial authorities, the Lab, and suptech 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

- **Superintendencia de Banca, Seguros y AFP of Peru (SBS):** Financial authority that regulates and supervises financial institutions, insurance companies, and private pension funds administrators in Peru, ensuring consumer protection 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

- **Constraints with prior social media analysis:** The social media analysis efforts previously undertaken by SBS were limited in terms of the number of sources that could be scraped and the depth of analysis. The solution primarily relied on sentiment classification and basic topic categorisation, which was manually performed by a local vendor. 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 tool. Moreover, the lack of integration between social media and other sources such as customer complaints required additional time and manual effort 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 and tools:** The tool could only be configured by the vendor, creating challenges concerning flexibility and control. SBS could not update classification, processing methods, alerts, or detection rules for potential market misconduct. This limited SBS's ability to customise and improve its supervisory tools to better meet specific needs.

Key features

- **Web and social media scraping and data integration:** Automates data collection from various web and social media platforms 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 to interprets opinions, sentiments, and emotions from text data, classifying them into positive, negative, or neutral categories.

- **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. models identify underlying themes in text data, helping to organise and understand large information sets.

- **Interactive dashboards:** Visualises data and insights with the filtering and drill-down capabilities to support supervisory investigations.

Benefits

- **Broad, automated social media monitoring and analysis:** The prototype developed automates data collection from various data sources, integrates it with other datasets, and applies real-time, advanced analytics to provide better insights into financial institutions' market conduct and consumers' behaviours, preferences, and sentiment.

- **Prototype availability:** The prototype remains fully available to SBS and customisable to meet its needs.

- **Reliable, enhanced supervision:** The prototype demonstrated SBS capabilities to shift from reactive to proactive supervision through an early warning system that identifies unusual patterns of transactions based on the analysis of customers' sentiments

[1] See Cambridge SupTech Lab (2024a), Financial Consumer Protection Suite with Web Scraping and Machine Learning- Based Analysis, Case Study N.1, August 2024, Cambridge: Cambridge Centre for Alternative Finance (CCAF), University of Cambridge. Available at https://lab.ccaf.io/wp-content/2024-Phiippines-casestudy

1. BACKGROUND AND SUPERVISORY CHALLENGES

In 2017, 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 and risk mitigation practices, treat clients fairly, comply with relevant regulations, and prevent unfair or abusive behaviors 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. These sources can be categorised as direct, such as customer complaints, and indirect, such as data posted on public, web-based platforms.

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 analysed 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 2020, SBS engaged a local vendor to scrape and analyse social media data to identify potential misconducts, including the inappropriate application of commissions, fees, and charges, misleading marketing practices, etc. While this solution captured relevant information using algorithms and hashtags linked to market conduct issues, its scraping, data preparation, and analytical capabilities were limited. For example, sentiment classification (neutral, positive, or negative) was not nuanced, and basic topic classification needed to be analysed manually. Moreover, the data generated by the local vendor could not be automatically integrated with other direct 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 and classification 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.

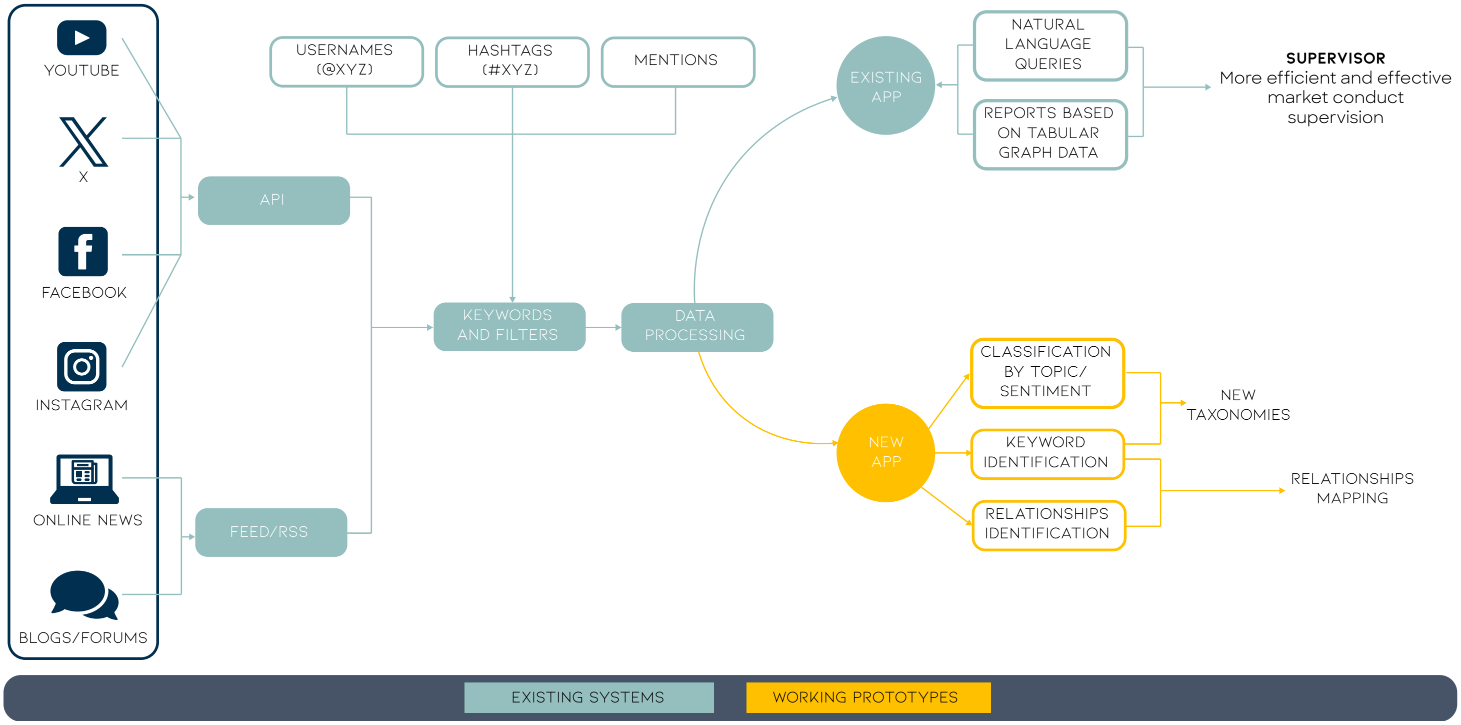
2. PROJECT CONCEPTUALISATION AND INCEPTION

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: (i) Relevant experience (60%), (ii) technical and managerial expertise (30%), and (iii) 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

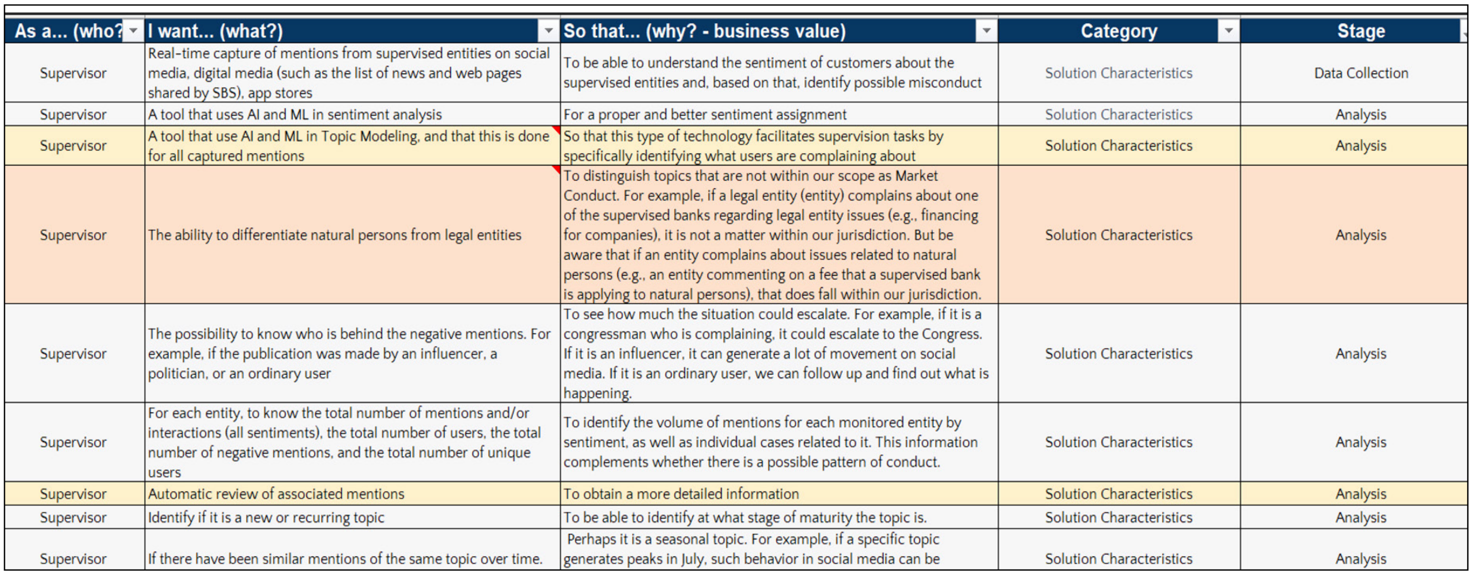
The project kicked off in June 2023, with the Lab facilitating the stakeholder and communication plan while leading design sprints with the cross-functional project team. FNA captured design documentation through user stories (Figure 2), technical requirements for desired features and functionality, 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 Lab's approach fosters a transparent environment where progress and details of development are visible to the financial authorities. This ensures that the system or solution being developed does not become a "black box" – a system or application whose inner workings are hidden or poorly understood by its users. By integrating user feedback throughout the development phases, the Lab's agile methods help evolve the solution to align closely with user needs and expectations, making the entire process open and understandable.

The Lab's approach fosters a transparent environment where progress and details of development are visible to the financial authorities. This ensures that the system or solution being developed does not become a "black box" – a system or application whose inner workings are hidden or poorly understood by its users. By integrating user feedback throughout the development phases, the Lab's agile methods help evolve the solution to align closely with user needs and expectations, making the entire process open and understandable.

The working prototype focused on utilising AI to classify social media posts by sentiment and topic, and keyword identification, which could inform new taxonomies and relationship identification and mapping. This data was then analysed and correlated with internal datasets, providing valuable insights into market conduct. The system enabled real-time monitoring and alerts, facilitating proactive supervision.

FIGURE 2. USER STORIES



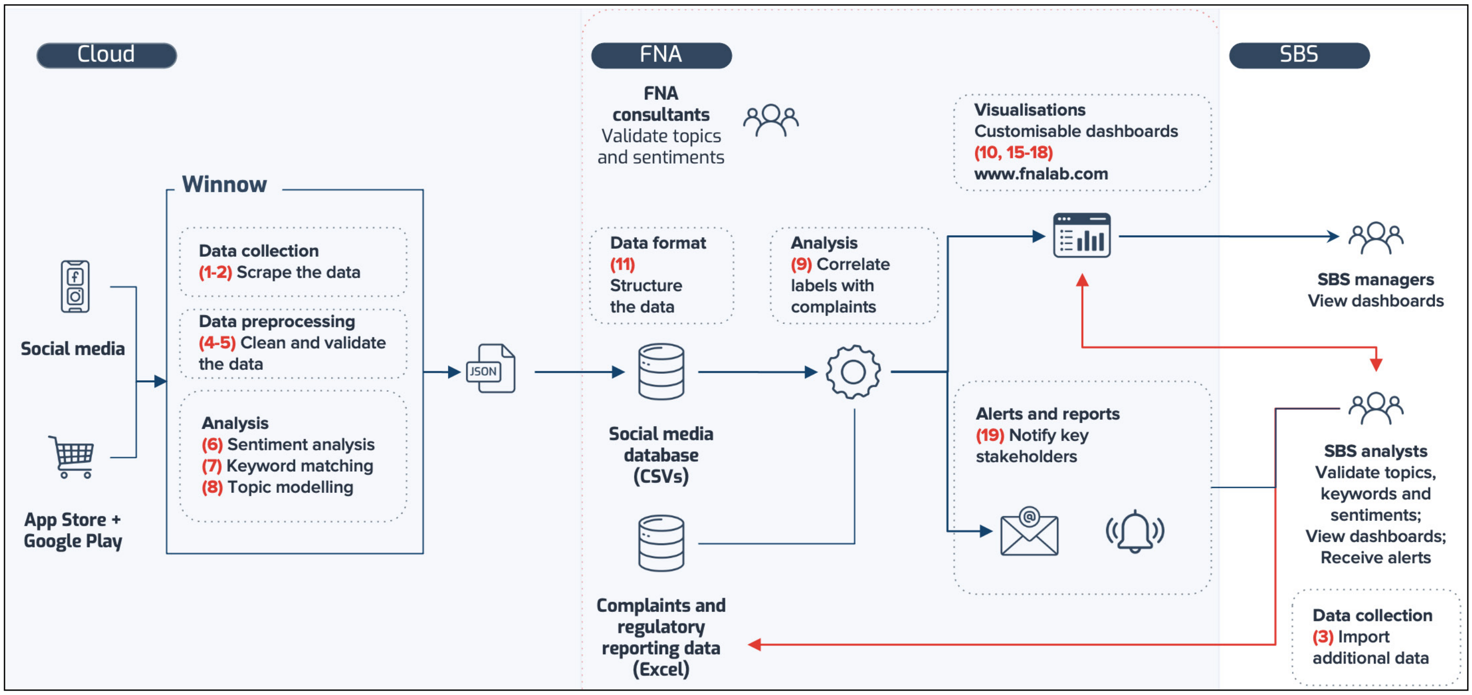
As a... (who?	I want... (what?)	So that... (why? - business value)	Category	Stage
Supervisor	Real-time capture of mentions from supervised entities on social media, digital media (such as the list of news and web pages shared by SBS), app stores	To be able to understand the sentiment of customers about 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

During the prototyping phase, it became evident that layering analytics on the pre-existing web scraping tool would not address the challenges. Social media sites often update their algorithms to detect and block scraping of their data thus identifying the need for a more advanced web scraping tool was required.

Another benefit of prototyping identified the challenges of extracting meaningful insights out of social media. The link between social media posts and market misconduct is opaque. Negative social media posts often contain sparse information, foul language, and irrelevant content, making it difficult for humans and machines to infer whether a complaint relates to actual market misconduct.

Any social media monitoring project of this kind should allow ample time for iterative improvements to the natural language processing models. Iterations over the models will require calibration on an ongoing basis as the environment changes.

FIGURE 3. SBS/FNA WORKING PROTOTYPE ARCHITECTURE DIAGRAM



5. THE APPLICATION

This project enhanced the scraping of social media, web, and other digital channels, and developed new analytics to further strengthen SBS's market conduct supervision capabilities.

Recognising the limitations of the existing social media data provider contracted by SBS, FNA conducted research to identify a vendor that could better address these challenges. They found that while several providers offer social media scraping services, most face limitations related to specific platforms, data volume, and content access. After their evaluation, FNA concluded that Winnow – the same vendor that had participated in delivering two other prototypes commissioned by the University of Cambridge in 2023[2-3] – was best positioned to scrape the necessary data to meet the project's objectives.

Various AI methods were combined to develop the analytics models, using real-time and historical data, to achieve the objectives of the working prototype. FNA found that a mix of BERT, Gensim (an open-source library for unsupervised topic modeling, document indexing, retrieval by similarity, and other NLP functionalities using modern statistical ML), and GPT (Generative pre-trained transformer, a type of artificial intelligence language model) together with classification and clustering ML models provided the most useful outcome for SBS. Results were presented in a cloud-based graph database with filtering and visualisation capabilities, including interactive timelines, tree maps, bar charts, and tables.

Three forms of advanced analytics were developed: sentiment analysis, topic classification and advanced detailed topic extraction:

- **Sentiment analysis** used NLPtown/BERT-based, multilingual model, fine-tuned for sentiment analysis. This is a supervised NLP model that was trained on approximately 50,000 labeled product reviews in Spanish. Each post was assigned a sentiment based on its contents; Positive, Negative, or Neutral.

- **Keyword tagging**: SBS specialists provided sets of keywords and categories that they want to group the contents into. The Natural Language Toolkit (NLTK) was used to clean the text for machine readability, including normalising and stemming text. Regular Expression (RegEx) was used to match keywords in the content to an expanded list of keywords for each category.

- **Detailed topic extraction:** Due to the absence of labeled data, an unsupervised topic modeling approach Latent Dirichlet Allocation (LDA) was employed to uncover latent themes within the corpus. To generate high-quality training data for the next stage, a human-in-the-loop process was implemented for selecting the topics of interest and sampling data for manual label assignment. This labeling process was iterative, ensuring the identified topics aligned with the specialised interests of the reviewers and domain relevance. To generate detailed topics, the labeled training data was used to run a supervised on the entire set of posts. For summarising, the BERT Extractive Summarizer was used to pull key sentences from the posts under each topic for a clear overview of the content. As an experiment, GPT was used to summarise contents into distinct bullet points from the summaries for easy readability.

[2] See Cambridge SupTech Lab (2024a), ibid.

[3] 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.

An ensemble of classifiers assigned a topic to each social media post and calculated the number and percentages of posts per topic. Summary bullet points (or subtopics) were formulated for each topic using (i) BERT Summarizer to extract the relevant sentences from the contents of posts about each topic and (ii) GPT to generate summary bullet points (or subtopics) in natural language based on those sentences.

Two possible solutions that perform the same underlying task of reclassifying posts using a more granular set of topic labels were explored. Both methods required manual validation to assess and iterate to improve the results.

Solution 1: Match bullet points back to posts

In the first solution, the summary bullet points remained unchanged and matched back to the original posts. There are two approaches to doing so:

A. Match relevant sentences (Before GPT)

The BERT Summarizer combines and shortens the content of all posts within a specific topic by removing irrelevant and repeated content. It does not generate new text, so for each relevant sentence, exact matches in the contents of one or multiple original posts can be found. Note that this process may not work in reverse; original posts that only contain irrelevant or repeated content cannot necessarily be mapped to one of the relevant sentences.

The tool could map relevant sentences back to the posts and assign an index to each of the relevant sentences before applying GPT. When using GPT to generate the summary bullet points the tool would modify the prompt to ask it to also provide, for each bullet point, the indices of the relevant sentence(s) that were used to generate it.

This method would track exactly those original posts and relevant sentences that lead to the summary bullet points and get an accurate match. However, some original posts may get missed, and it was unclear if GPT is fully capable of prividing a sensible output given the modified prompt.

B. Match summary bullet points (After GPT)

The output of GPT could be matched directly back to the original posts using a combination of similarity matching and regular expression. This could be implemented and tested to matching more quickly. However, because GPT reformulates and shortens relevant sentences, matching bullet points to original posts will produce more error. So, several iterations would be needed to lower the error rate to a satisfactory level, which could end up taking longer than implementing Solution 1A.

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

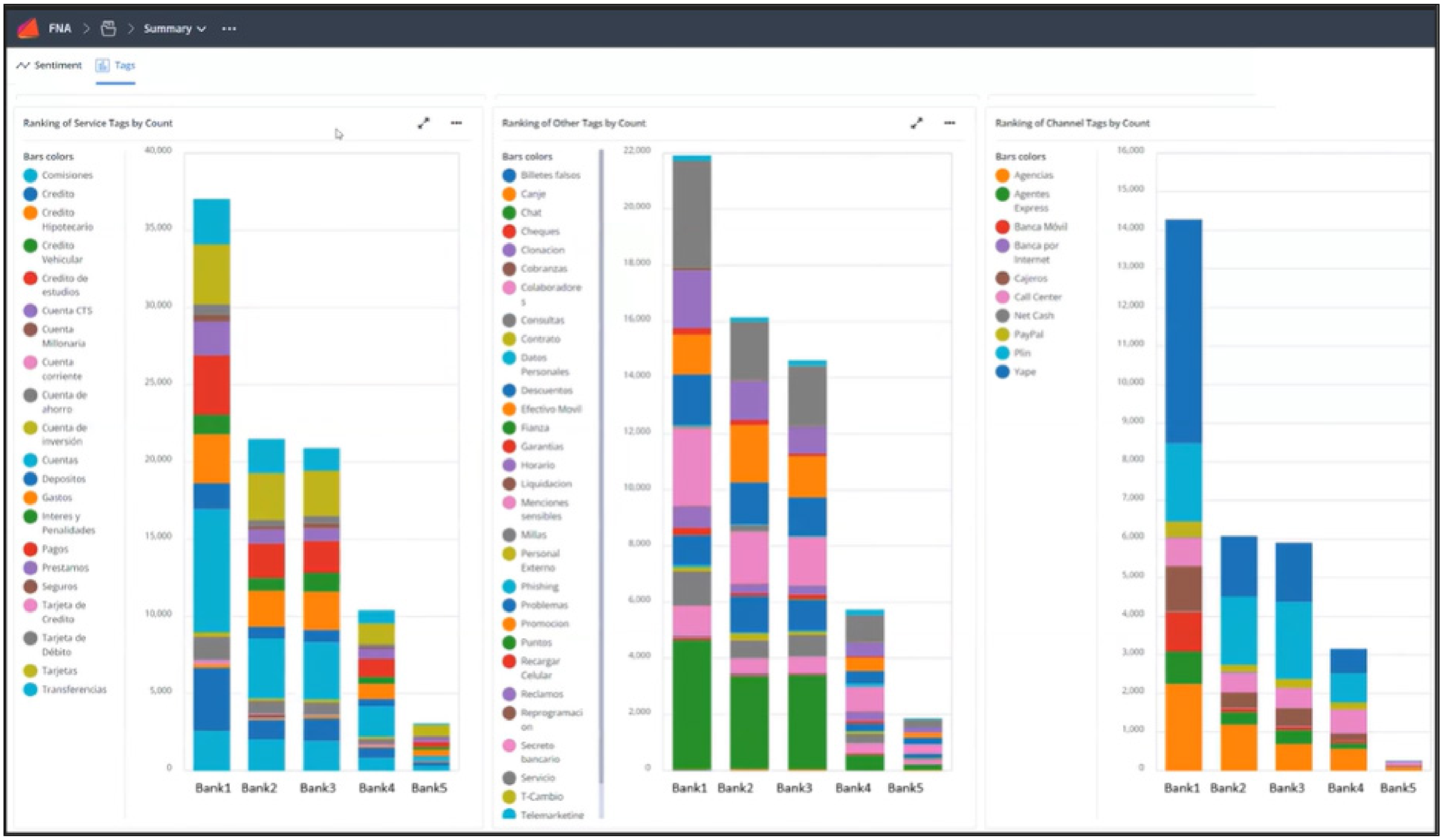
A more stable, long-term solution is to drop the BERT Summarizer and GPT steps entirely and create new, granular subtopics instead of the summary bullet points. A method similar to the one for creating the topics could be used, that is, a mix of unsupervised learning to identify clusters, manual review of a sample of posts to assign sensible subtopics, and supervised learning to assign subtopics to the remaining population of posts.

A subject matter expert at SBS would work with FNA to validate the subtopics and provide good labels for a large sample of posts to be used as the training dataset. This would provide SBS more flexibility in defining the subtopics with validation built directly into the workflow.

This solution required more manual work at the beginning of the process and occasional review efforts to ensure that the subtopics are in line with what is discussed on social media and app stores.

The Summary Dashboard (Figure 4) displays 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 provided by SBS for monitoring and analysis, including call centres, branches or payment platforms.

FIGURE 4. FNA'S SUMMARY DASHBOARD



The Sentiment Analysis Dashboard (Figure 5) displays data scraped from social media for five large 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 (Figure 6) 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. that warrant further investigation.

FIGURE 5. SENTIMENT ANALYSIS DASHBOARD

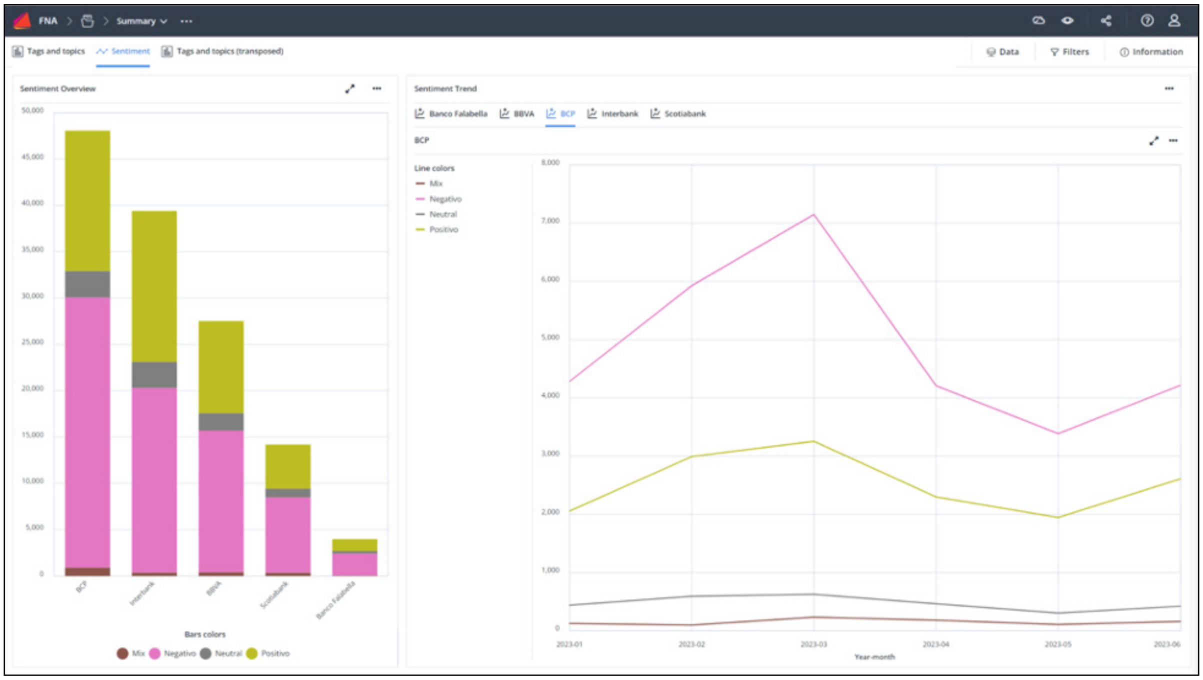
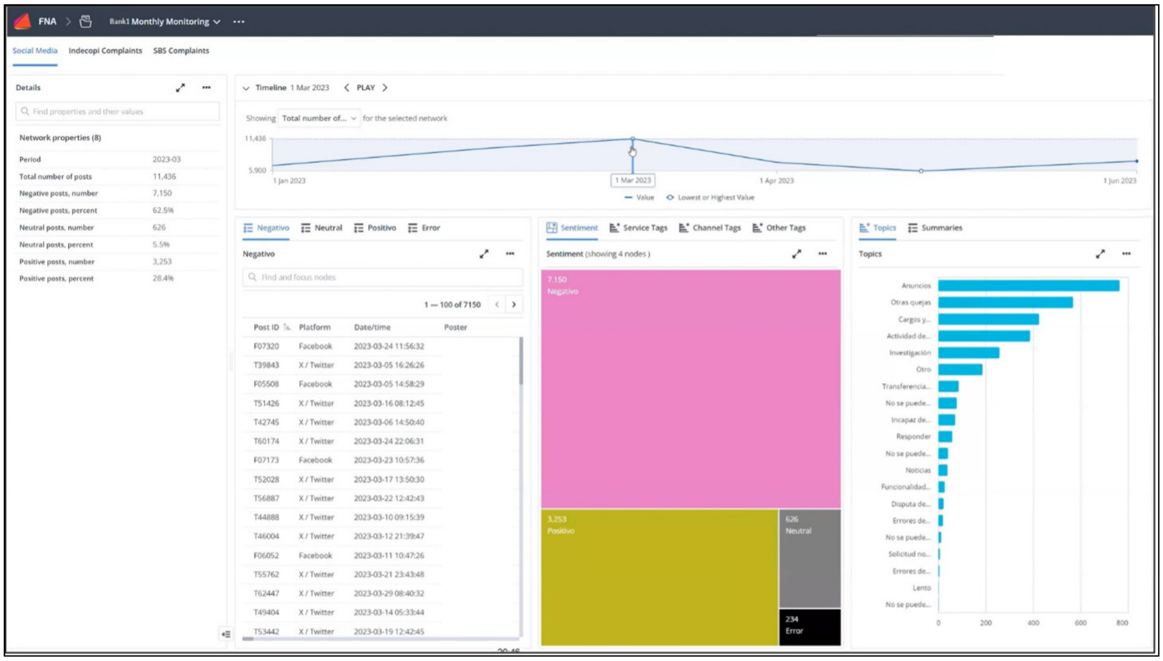


FIGURE 6. FNA'S MONTHLY MONITORING DASHBOARD FOR INDIVIDUAL BANKS

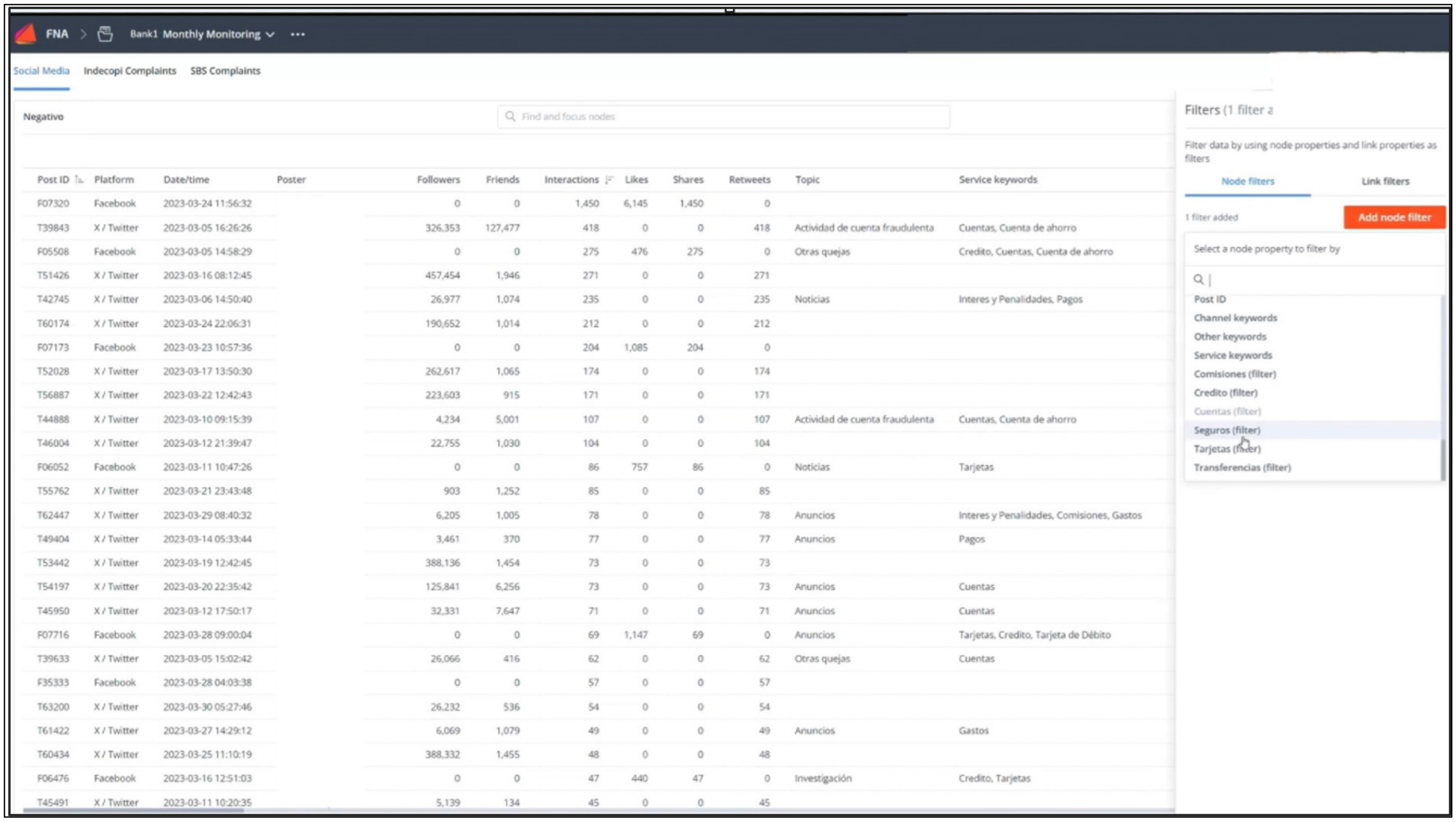


Network properties (8)

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%

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 7. FNA'S FILTERABLE POST DASHBOARD



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		

Filters (1 filter a

Filter data by using node properties and link properties as filters

1 filter added

Select a node property to filter by

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.

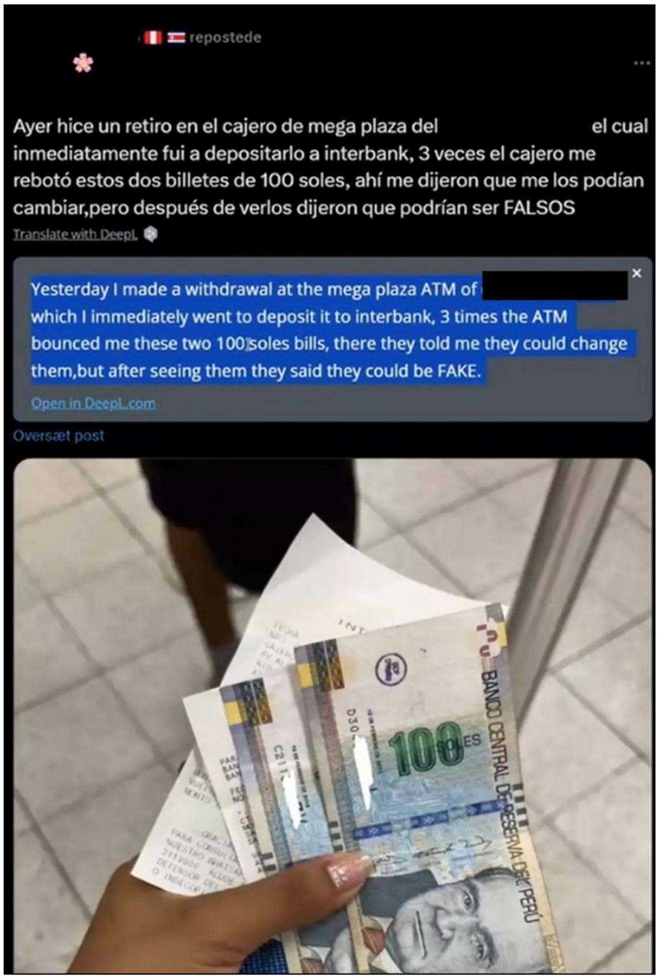
This dashboard can be filtered by specific keywords, and be sorted by the level of interactions, such as the total number of reposts, likes, and shares. In this example, a news agency's post ranks highest on X (formerly Twitter). The second most interacted-with post comes from an individual user.

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

FNA developed three major and eight minor prototype versions via quality iterations with each new version enhancing the data filters, topic modeling and user interface of the previous version. The result included 53 widgets on 20 pages across 8 dashboards, allowing users to analyse hundreds of thousands of social media posts efficiently.

These dashboards allowed SBS to visualise 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 8. SAMPLE POST



repostede

Ayer hice un retiro en el cajero de mega plaza del [redacted] el cual inmediatamente fui a depositarlo a interbank, 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 interbank, 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 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
- 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 of social media and web-based platforms to report real-time information about market conditions and emerging consumer risks. This allows agencies to make more informed decisions and manage risk effectively.
- **Bridge the gap between reactive response and preventive action** through proactive monitoring, ensuring potential risks are identified, analysed, and addressed before they escalate.
- **Receive near real-time updates** on public sentiment 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.

7. WHAT´S NEXT

SBS confirmed that the solution accelerates their digital transformation by clearly demonstrating the value of integrating artificial intelligence into the supervisory process. The prototype exercise helped them understand the steps for moving into production. The Lab's approach is specifically designed to avoid vendor lock-in, ensuring that SBS retains flexibility. They will continue to have access to the working prototype environment and a perpetual license to both the object and source code of the solution.

As part of the Lab's Application Incubation programme, 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.

> *“The data and insights gathered through social media monitoring enabled by the prototype offer a versatile resource for enhancing our policy development process at SBS. Firstly, the tool allows us to identify areas where new policies are needed proactively. For instance, if there's a spike in consumer complaints on a specific issue, it enables us to recognise the need for a targeted policy to address and mitigate that concern. Furthermore, such a solution plays a key role in evaluating the impact of existing policies. By collecting mentions of compliance with our current policies from various entities, we can assess their effectiveness and determine if adjustments are necessary. This ensures that our policies remain robust and adaptable to changing circumstances.*
>
> *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 policies, 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*

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