Reaching the forefront of Gen AI in banking – by decoding "Voice of customer" on social media interactions

When Gen AI emerged as the next leap in technology in early 2023, organizations faced pressing questions on its most effective applications. Many of them have grappled with how to integrate and scale the technology, and the potential business value impact of its use cases.

Within banking industry, while initial Gen AI pilots and efforts focused on productivity, recent challenges and use cases are more focused on answering how this technology could alter customers interactions with the banks, and the execution of various job functions.

A recent study by McKinsey Global Institute highlights that the banking sector stands to gain significantly from the adoption and use of Generative AI, with potential value creation estimated at $200-340 billion annually. From this, ~20-30% of the gains are projected to come from retail and corporate banking each.

The potential value creation from the adoption of Generative AI is estimated to be $200-$340 billion annually

Strategically implementing customized Gen AI solutions allows financial institutions to significantly elevate the overall customer experience and improve service operations.

One key area on which our team has been collaborating with various banking clients, is to leverage Gen AI to provide better customer assistance, especially on social media touchpoints.
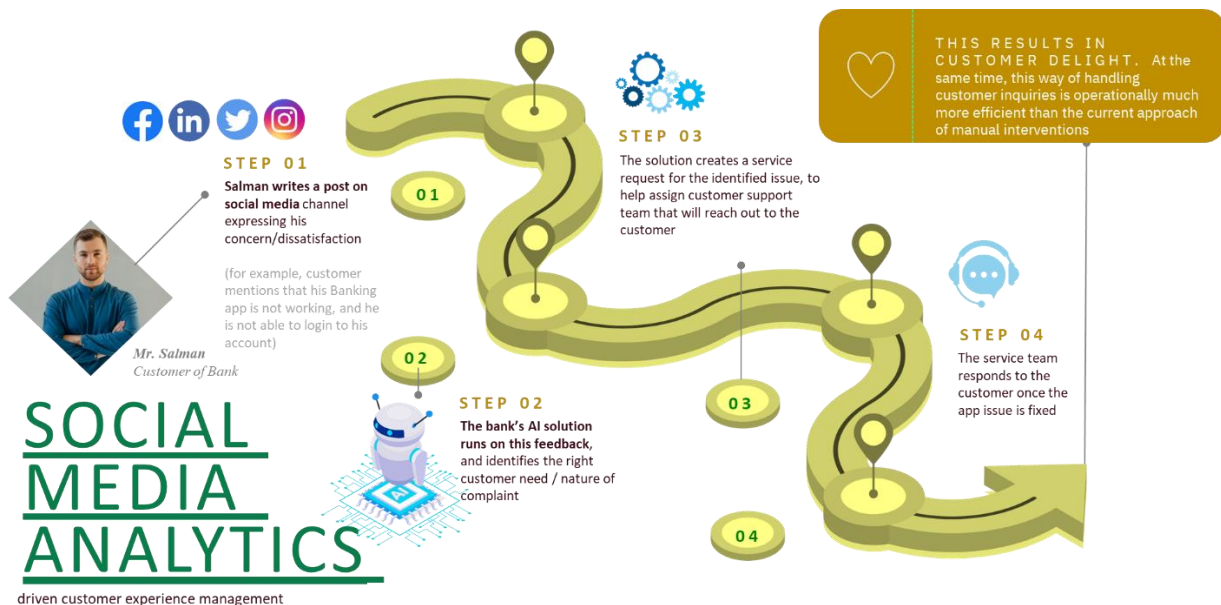
While 40-45% banking customers have traditionally voiced their issues using channels like chatbots on mobile app/online banking, or by speaking to a live agent (~20%), they are increasingly turning to social media platforms, online reviews, and customer service mails, to raise their voice more openly and intensely.

Coupled with this is the fact that banks have generally been behind the curve in terms of developing effective technology solutions or bespoke tools, to manage social media interactions: Some of the key challenges that have persisted in this area include:

1. **Longer wait times**: The bank may prioritize complaints or queries raised through traditional channels first, potentially leading to extended wait times for customers using alternative platforms
2. **One way communication:** Social media interactions have often been a one-way street, with many banks lacking a strong technological foundation to prioritize and respond to these interactions promptly
3. **Brand reputation risks:** Public tweets and reviews on platforms such as Google can have a significantly negative impact on brand perception, influencing the opinions of other customers
4. **Missed opportunities in satisfaction assessment:** while banks have continued to rely on customer surveys for NPS and CSAT, there is untapped potential in leveraging Gen AI applications, to truly understand the "voice of customer" from their feedback

Given this context, our engagement aimed at helping the bank design and implement a gen AI solution, which could identify the key pain areas using customer interactions and feedback on social media, and recommend the right action automatically to solve the issue.

SOCIAL MEDIA ANALYTICS

driven customer experience management

Mr. Salman
Customer of Bank

STEP 01
Salman writes a post on social media channel expressing his concern/dissatisfaction

(for example, customer mentions that his Banking app is not working, and he is not able to login to his account)

STEP 02
The bank's AI solution runs on this feedback, and identifies the right customer need / nature of complaint

STEP 03
The solution creates a service request for the identified issue, to help assign customer support team that will reach out to the customer

STEP 04
The service team responds to the customer once the app issue is fixed

THIS RESULTS IN CUSTOMER DELIGHT. At the same time, this way of handling customer inquiries is operationally much more efficient than the current approach of manual interventions

However, implementing generative AI comes with certain risks. It cannot be deployed as an out-of-the-box solution; it requires careful customization to ensure the responses are contextually relevant. Additionally, clear guardrails must be established to mitigate communication risks and ensure that the AI operates within defined boundaries.

> **Implementing generative AI comes with certain risks. It cannot be deployed as an out-of-the-box solution**

We have addressed these challenges by carefully embedding personalization into our Gen AI solution. The model and aspect identification have been tailored to align with the bank's specific data and downstream communication plan.

Technical solution approach

Further below are details of our technical solution approach and the model architecture:

1. **Data collection** – historical data for all customer complaints, sales inquiries and queries, from various social media platforms such as X, Facebook, LinkedIn, Instagram etc.
2. **Data pre-processing** –
   a. Data cleansing and tokenization (using a pretrained transformer tokenizer)
   b. Text encoding and batch creation with random sampling
   c. Applying language detection models (such as NLLB200 from Meta) for translating non-English languages such as Arabic into English
3. **Intent recognition** – training and customizing a LLM to identify sales leads and complaints from the customer interaction texts. *The details of the modeling approach is present in the subsequent section*
4. **Categorization and aspect extraction** – extract the specific product and service-related context for which the customer is reaching out
5. **Output generation for downstream integration** – exposing the model output to seamlessly integrate with the bank's systems, either by creating service request tickets within the complaint management system or by generating sales leads for CRM integration

Model architecture:

Our team trained and customized a BERT (Bidirectional Encoder Representations from Transformers) model, to identify intent of customer interaction text.

- **Bidirectional:** Considers context from both left and right sides of the target word
- **Pre-trained:** Trained on a large corpus to capture diverse language patterns
- **Transformers:** Utilizes transformer architecture for efficient processing
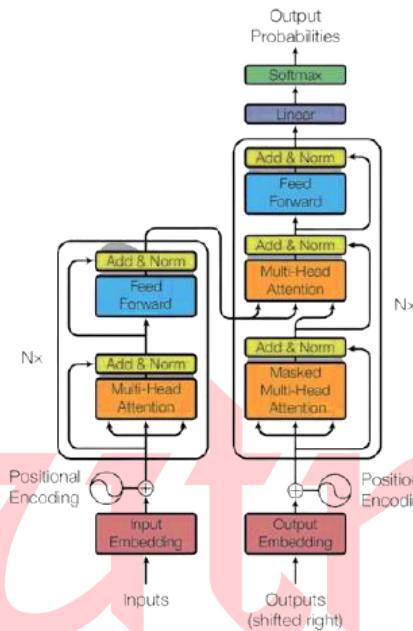
Some of the key model components were:

- **Encoder:** Multi-layer transformer blocks that process input text
- **Input Representations:** Combines token, segment, and position embeddings
- **Output Representations:** High-level representations for downstream tasks

The model was trained using a relatively small sample of ~10,000 historical data points (since a pre-trained model doesn't require a large no. of data points), by using 12GB CUDA enabled GPU system.

Model training approach:

Optimizer Initialization

STEP A
- Create AdamW neural network optimizer with learning rate = 10^{-5}
- AdamW optimization is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments with an added method to decay weights

Forward Propagation

STEP C
- Pass one mini batch at a time into transformer model and get the output
- Compute the loss from expected output

Initialize Transformer model

STEP B
- Initialize pretrained model weights in transformer model(bert-base-uncased)
- Add one dense layer at the end of the model as the output layer with n neurons, where n is the number of classes in the dataset

Back Propagation

STEP D
- Propagate the loss backward
- Use the optimizer to compute the delta change in each weights and update the weights to minimize the loss

Model Evaluation and Retraining

STEP E
- Evaluate the model performance on validation dataset
- Compute the Precision, recall and F1 score corresponding to each labels
- Reiterate the process with the complete dataset for n epochs (initially 10)

Model bootstrapping and hyperparameter tuning:

For bootstrapping, team used the following steps:

1. **Model inference** – applying model on unlabeled dataset and getting prediction probabilities
2. **Filtering uncertain predictions** and expanding dataset by adding new predictions
3. Model **fine-tuning** based on new dataset
4. Evaluating performance on validation and test data-sets, including cross-validation
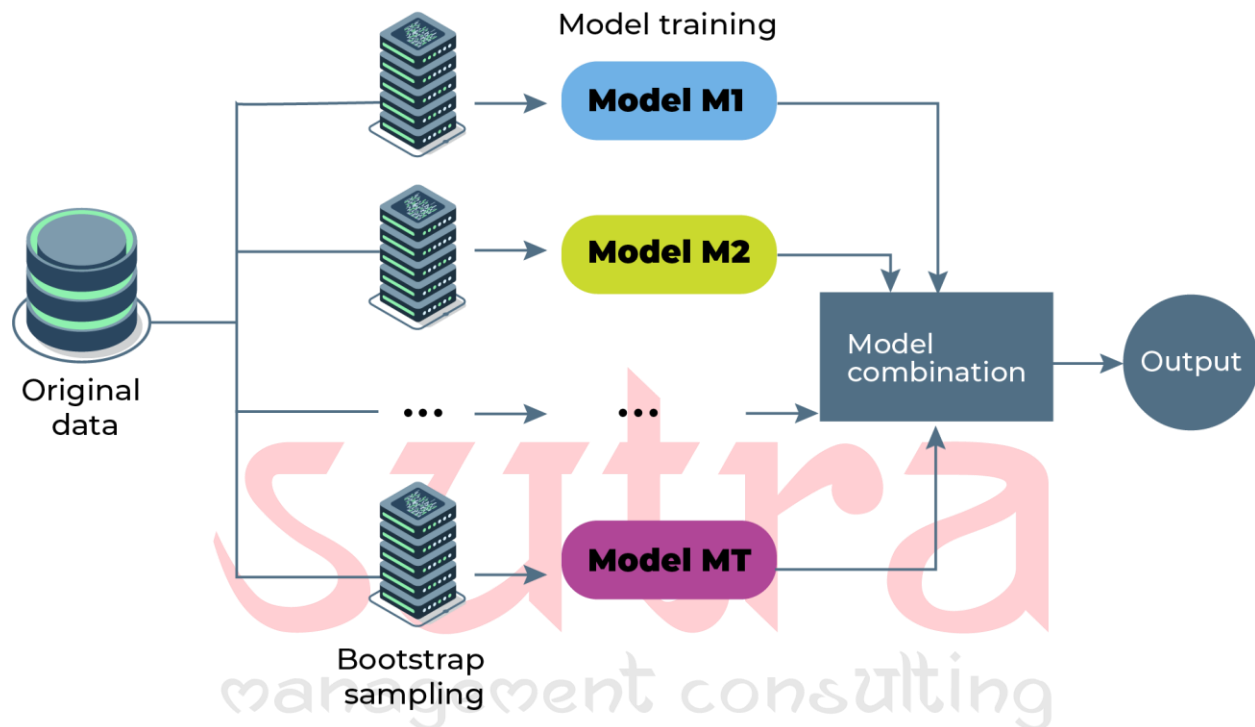
Hyper-parameter tuning:

We performed hyperparameter tuning using the following approach:

1. Iterations based on Optimizers (AdamW, AdaFactor), Batch size (4,8,16), # of epochs (10, 20, 30, 40, 50), and learning rate (10^{-3}, 10^{-5})
2. Grid search for the iterations
3. Evaluate performance across precision, accuracy, F1 score, recall etc.
4. Choose the best performing hyper-parameters
5. Fine-tune the BERT model on full training dataset

For topic extraction, our team worked with two approaches:

TF-IDF analysis:

Data Collection

Utilize a prepared dataset during the model training stage for intent classification

N-gram Generation

- Generate n-grams (sequences of n consecutive words) from the pre-processed text data.
- Include unigrams (single words), bigrams (pairs of consecutive words), and trigrams (triplets of consecutive words)
- Higher-order n-grams may capture more complex relationships but can also increase computational complexity.

TF-IDF Calculation

- Compute the Term Frequency-Inverse Document Frequency (TF-IDF) scores for each n-gram.
- TF measures the frequency of a term in a document, while IDF measures the importance of a term across multiple documents.

Keyword Extraction

- Sort the n-grams based on their TF-IDF scores in descending order. Select the top-ranking n-grams as keywords for each social media post.
- The number of keywords to select can be predetermined or based on a threshold TF-IDF score

Post-Processing

- Remove any n-grams that are irrelevant or not suitable as keywords (e.g., common phrases, generic terms).
- Normalize the selected keywords (e.g., convert to lowercase, lemmatize) for consistency

Latent Dirichlet allocation (LDA):

Data Collection

Utilize a prepared dataset during the model training stage for intent classification

N-gram Generation

- Generate n-grams (sequences of n consecutive words) from the pre-processed text data.
- Include unigrams (single words), bigrams (pairs of consecutive words), and trigrams (triplets of consecutive words)
- Higher-order n-grams may capture more complex relationships but can also increase computational complexity.

Topic Modelling

- Apply LDA to the pre-processed text data to identify latent topics. Determine the optimal number of topics based on coherence scores
- TF measures the frequency of a term in a document, while IDF measures the importance of a term across multiple documents.

Keyword Extraction

- Combine the results from n-gram analysis and topic modelling to extract keywords.
- Assign IDF weights to terms based on their relevance to each identified topic. Terms with higher weights within specific topics are considered keywords related to those topics

Post-Processing

- Remove any n-grams that are irrelevant or not suitable as keywords (e.g., common phrases, generic terms).
- Normalize the selected keywords (e.g., convert to lowercase, lemmatize) for consistency

Impact and outcome:

For 20% of complaints
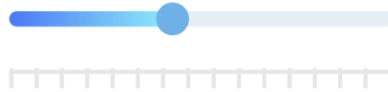
processing time has reduced from **24 hours** to less than **15 minutes**

+7% agents efficiency

Customer service agents improved their operational efficiency in logging and managing service tickets and leads
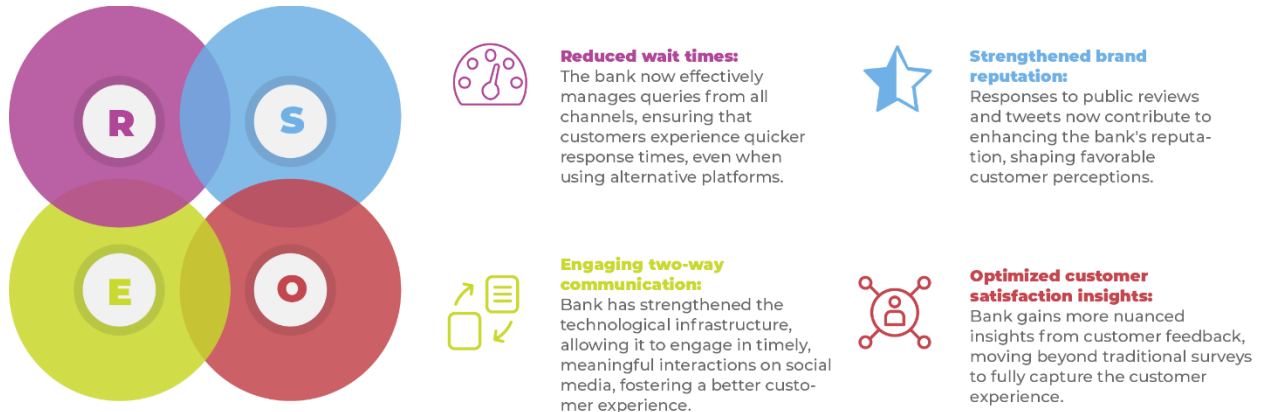
Immediate severe escalation resolution —10%

In about 10% of severe escalation cases, they were able to provide an instant resolution

Our team was able to create significant value for our banking clients in 3 key areas:

1. **Improved sales potential**: Within first few months of implementation, the solution identified several hundred incremental sales leads for cards and loans each month, from prospects expressing interest on social media
2. **Enhanced customer experience**: The bank's customer service team reduced the average response time for nearly 20% of complaints from 24 hours to less than 15 minutes. In about 10% of severe escalation cases, they were able to provide an instant resolution

Category
COMPLAINT
Issue
FEE RELATED ISSUE
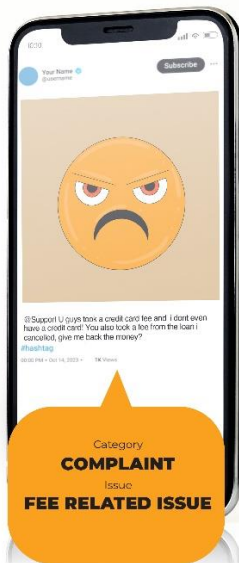
3. **Agent performance**: With clearer insights into customer intent and context, customer service agents improved their operational efficiency in logging and managing service tickets and leads by 7-8%

Reduced wait times:
The bank now effectively manages queries from all channels, ensuring that customers experience quicker response times, even when using alternative platforms.

Strengthened brand reputation:
Responses to public reviews and tweets now contribute to enhancing the bank's reputation, shaping favorable customer perceptions.

Engaging two-way communication:
Bank has strengthened the technological infrastructure, allowing it to engage in timely, meaningful interactions on social media, fostering a better customer experience.

Optimized customer satisfaction insights:
Bank gains more nuanced insights from customer feedback, moving beyond traditional surveys to fully capture the customer experience.

As the solution continues to evolve, we are collaborating with our banking clients in using gen AI for further enhancing customer experience by:

1. Generating insights from customer feedback and **ratings given on Google** for specific branches, products and services
2. Instant trend detection for **monitoring shifts in customer opinion** and engagement patterns in real time, allowing the bank to quickly adjust its communication strategy
3. **Automated summarization:** Assisting bank's call center agents by automatically summarizing notes taken during customer interactions, streamlining the process of servicing calls
4. Tracking interactions: identifying **influential users** or **brand advocates**, allowing the bank to engage effectively with key stakeholders and manage its online presence

About the authors:

Saurabh Assat, Partner Sutra Dubai
Ravi Saroj, Technical Expert Sutra India
Rahul Bijarniya, **Senior** Consultant Sutra Delhi