




Data Search and Discovery in Banking

An Overview of AI Capabilities



A collaboration between Iron Mountain and Emerj



ARTIFICIAL INTELLIGENCE AND BANKING

An Introduction

Banks seem to be adopting AI applications ranging from business process automation to credit scoring. Historically, banks have collected vast amounts of data records and even some of the more conservative banks tend to have the resources needed for AI projects. Efficiently navigating the vast data stores to gain valuable business insights involves understanding the capabilities of AI in information search and discovery applications for the banking sector.

One of the earliest applications of text mining in banking was the development of Automatic Processing of Money Transfer Messages (ATRANS). This system could extract information in a predefined template from telex messages, which were sent between banks to confirm money transfers.

The telex messages were highly predictable in their content, and the standardized format made ATRANS a successful system. AI is not a magic fix-it-all solution for challenges faced by banks today. Rather, business leaders can think of it as a tool that can help them find ways to improve outcomes that already exist within their data. AI software is only as good as the data it consumes.

Businesses face many challenges in terms of collecting and structuring their data, which is key to the successful implementation of AI systems. An AI application is only as good as the data it consumes.

Natural language processing (NLP) and machine learning models often need to be trained on large volumes of data. Data scientists tweak these models to improve their accuracy. This is a process that might last several months from start to finish, even in cases where the model is being taught relatively rudimentary tasks, such as identifying semantic trends in a banking company's internal documentation.

Most AI systems necessarily require the data to be input into an AI system in a structured format. Businesses would need to collect, clean, and organize their data to meet these requirements. Although creating NLP and machine learning models to solve real-world business problems is by itself a challenging task, this process cannot be started without a plan for organizing and structuring enough data for these models to operate at reasonable accuracy levels.

Germán Sanchis-Trilles holds a PhD in Computer Science from the Polytechnic University of Valencia and co-founded Sciling. He has nine years of research experience in the fields of natural language processing and pattern recognition. Regarding data management, Sanchis-Trilles urges business leaders to understand the following:

DATA VELOCITY

What is the rate at which the company is receiving or producing data?

DATA STORAGE

How and where can the data be stored? What are the resources required for storing this data?

VERACITY AND SECURITY

Is the data accurate and secure? Is the data coherent?

Is the data regulation compliant?

In the banking sector, there are also other challenges with data in terms of security and access, privacy, regulations and compliance, IP protection, and physical and digital barriers.

Most NLP and text/data mining applications are developed based on existing software libraries and established use-cases. International banks that collect data in languages other than English might also find it difficult to develop NLP software that can directly understand contextual documents in local languages.

For instance, the fact that there are currently more NLP libraries for English than German might force a German bank to take the additional step of converting its documents to English before building any AI applications.

QUESTIONS TO ASK

to determine if your business's data is ready for use with AI

Is your data accurate? As mentioned at the start of this report, AI systems are largely dependent on the quality of the data on which they are trained. For example, if two customer's credit histories are swapped in a dataset, a predictive model trained on that data might suggest inaccurate credit scores for those customers.

Do you have the right kind of data? The data must be relevant to the objective of the AI system, and the data necessary to achieve the objective must be present in the datasets that are fed to the AI model. Credit histories are integral to the prior example. Existing loan and social media information might also inform the credit scores of the two customers.

Is all of your historical data consistent? Businesses need to ensure that their data is consistently collected in the same format. Large banks operating out of many global locations might need to ensure that their offices are collecting data in the same formats, which could be challenging.

Does your data need to be updated? The data being fed into an AI system changes in most use cases, and the system may need to be updated. Businesses need to update their data regularly to ensure accurate insights are being derived from their AI system. A machine learning model trained on old data might provide less relevant or wholly inaccurate outputs than one updated frequently with new data. Credit scoring and mortgage management software requires up-to-date customer data, else the software might make predictions that do not accurately reflect a customer's financial situation. This could lead to ill-informed underwriting and investing on the part of the bank.

AI FOR SEARCH AND DISCOVERY

[A report from the Financial Security Bureau](#) seems to suggest that NLP and machine learning can be applied to several applications in banking, such as customer-facing banking chatbots and credit scoring.

Data search and information discovery usually involves several steps even after the data has been organized and collected. NLP and machine learning can be used to create a searchable index of all internal documents. After indexing, the data can be made searchable through an interface. Search interfaces could offer basic search options, such as Boolean (and/or/not), segment, numeric range, or advanced search options that might include natural language search, fuzzy search, and concept search.

The same tools driving advances in machine learning in search engines are being adopted in the banking industry. For example, entity recognition tools similar to those used in search engines are now being used to identify news and social media conversations relevant to publicly-traded firms. As more firms adopt NLP and machine learning methods, the incentives to access additional data may increase among banks.

Giacomo Domeniconi, a post-doctoral researcher at IBM Watson TJ Research Center and adjunct professor for the course “High Performance Machine Learning” at New York University, suggests that search tools might get even better in the near future:

“Search tools which can contextually retrieve information from both structured and unstructured data might not be that far away from now. This might especially be true in sectors like banking where the companies have the economical resources to spend on gathering information from both structured and unstructured data.”

USE-CASES Applications of AI in banking

USE-CASE 1 Mortgages

According to the AI and data science experts at Iron Mountain, data search and discovery can help banks automate certain operations in mortgage management. Evidence suggests that banks are engaging in structured data capture through forms (HUD-1, tax forms, and loan applications). That said, unstructured documents such as PDFs, audio files, video files, and hand written notes still account for almost 80% of all documents in the mortgage industry, according to Axis Technical.

Banks can use AI to extract information from structured, semi-structured, and unstructured documents. Structured documents are usually easier to prepare for input into a machine learning software. Unstructured documents can be digitized using advanced optical character recognition (OCR) or computer vision. A bank can then apply NLP techniques to the extracted textual data and gain insights from all the different types of data they are collecting.

Semi-structured documents, such as invoices and closing statements, have parts that are structured and others that may even be handwritten. AI software can help digitize and organize the information that a bank might consider critical from these documents.

Unstructured documents present the biggest challenge for document imaging since the metadata that banks need to extract in order to train machine learning models is free-form and might be a sentence, paragraph, or whole page within a document.

Radim Rehurek, who earned his PhD in Computer Science from the Masaryk University Brno and founded RARE Technologies, points out:

“Historically a lot of the internal workflows in banking or insurance, deal with manual data entry in some form. A significant portion of these internal documents are also usually unstructured documents such as PDFs, hand written notes, audio or video files.”

Automated document processing using AI software in the mortgage sector might be a go-to low-hanging fruit application. Companies in the mortgage sector will find that this automation might help them reduce human labor costs, including any data-entry training costs.

NLP and machine learning can also help banks speed up loan and mortgage application processing, especially for large banks with volumes of incoming forms. The scale makes using AI software highly lucrative as opposed to human officers that would need to manually review the documents. AI can help banks analyze mortgage and loan applications to see if the forms are missing any information and immediately send customers requests for the same. This could allow the reviewing officers to steer clear of applications that have rudimentary problems (such as missing information) and only deal with cases involving more complex issues.

AI software for automated data capture could also help banks reduce human error and operate at a scale that might not be possible for humans to compete. Once all the documents have been scanned and the data is

captured and organized, mortgage firms can create a front-end search interface that allows employees to have access to the information and documents they need to make strategic business decisions.

That being said, training these algorithms requires a large amount of data, and small or medium size mortgage operations with less than 100 transactions per week might find they may not meet the data requirements for training an AI model.

AI for information search and discovery in mortgages might be most useful for large mortgage operations with over 500 transactions per week. Such firms can potentially capture data from web pages, word documents, emails, and other forms of structured text to make it accessible to employees.

Another prominent application for AI-based data search and discovery in banking is in credit scoring for loans. The value of a loan is directly related to how likely a bank thinks an individual or a business may default on that loan. Credit scoring is a critical part of the loan process because it helps banks identify the likelihood of default by accessing customer data, including credit histories, social media posts, and, in some cases, an individual's entire digital footprint.

The vastness of this data makes it almost impossible for human analysts to comb through and identify customers with a high probability of defaulting. NLP and machine learning can help banks crawl through digital footprint data, such as social media posts, internet browsing data, geolocation, and other smartphone-captured information, to generate a credit score for each customer.

Most credit scoring models have historically been based on transaction and payment histories between customers and banks. But banks seem to be increasingly turning to unstructured and semi-structured data sources to capture a more nuanced view of creditworthiness and improve the accuracy of credit scores.

Several AI vendors, such as Lenddo and ZestFinance, already offer credit scoring solutions for banking customers. The value proposition of these vendors seems to be around speeding up lending decisions and processes while simultaneously limiting risks. That said, according to the Financial Security Board, it has not yet been proven that machine learning-based credit scoring models outperform traditional softwares drastically in assessing creditworthiness.

Banks have several channels for collecting data from customer interactions through social media conversations, emails, phone calls, and website form-fills. With large amounts of data being collected about banking customers, the other broad application for search and discovery AI seems to be in customizing products to improve customer engagement.

We interviewed Adnan Masood, who earned a PhD in Machine Learning from Nova Southeastern University and is Chief Architect of AI and Machine Learning at UST Global. Masood is also a visiting scholar at the Stanford Artificial Intelligence Lab and MIT AI Lab. Masood points out:

“AI-driven data search and discovery primarily helps understand the intent. The consumer might be looking for a specific fund, a credit card, loan, an ETF, or checking the health and progress of existing products. A cognitive search provides a richer experience by integrating likes and dislikes, life events, and transaction history, providing context-sensitive results for products which fit these individuals in their specific stages of life and needs.”

According to Masood, AI data search and discovery tools help map these recommendations for customized product development. Large banks are exploring the possibilities of “Segment-of-One” financial products, which involve the ability to track and understand individual customer behaviour).

In the banking industry, customer interactions and opinions are usually very clear and are rarely neutral. More often, customer reviews are highly positive or highly negative. Negative statements from customers are also usually

much lengthier than positive reviews. Large banks collect vast amounts of customer information from multiple channels and need a way to personalize the banking experience.

By analyzing conversations from chatbots, customer feedback forms, social media pages, and transaction histories, a bank might discover the customer’s intent faster and more accurately. For instance, [the Royal Bank of Scotland \(RBS\) claims they use NLP](#) text-mining techniques to extract trends from customer feedback. According to RBS, they deployed an AI software that could ingest data from customer emails, surveys, and call center conversations to identify which issues affect their customers the most.

Masood provides context with a personal anecdote:

“A large financial institution in Australia whom I work closely with has applied cognitive capabilities, including NLU/NLP, to provide categorical natural language search in a variety of spending categories.... The financial institutions which get organized around their data can use this information to invest in creating better experiences and returns for their clients.”

Banks can also effectively use search and discovery AI software to provide their customers with a complete view of their interactions with the banks when it comes to their accounts, share deposits, and insurance contracts.

With AI-based search and discovery, banks could discover where customers spend most of their money and open up potential cross-selling opportunities. The scale of the customer-facing side in the banking sector makes it almost impossible for a human team to sift through all the data.

AI-DRIVEN BUSINESS VALUE

Data search and discovery applications for AI in the banking sector are emerging rapidly due to the availability of vast amounts of data and large multinational companies with the resources to implement new use cases.

AI search and discovery might automate repetitive tasks for banks and improve sales and customer service processes on the front-end.

Among all the applications discussed in this paper, credit scoring seems to have the most traction. Several AI vendors offer credit scoring solutions to banks. AI has made it possible to gain a deeper understanding of the default

risks for customers by enabling banks and financial institutions to analyze their entire digital footprint to generate their credit scores.

This has also made it possible for human loan officers to search and retrieve the data relevant to a case from both structured and unstructured documents. As a result, loan officers can review applications faster and more accurately.

The availability of historical data across a range of bank's customers and products would determine the success of such AI implementations.

About Emerj

Emerj is where business leaders turn to understand how AI is impacting their business and industry and what to do about it. We're the industry source for authoritative market research and competitive intelligence for the business applications of artificial intelligence.

Our objective, jargon-free research and industry overviews are designed to give executives and decision-makers exactly what they need for competitive insight, informed AI technology procurement, and strategic planning around AI.

AI disruption will open up new vistas for opportunity, creating new business paradigms and making others extinct. Emerj is the best place for leaders to get the facts and strategic insight they need to stay on the winning end of AI's inevitable disruptions.

Emerj offers custom research and thought leadership for brands that are innovating in artificial intelligence in various sectors and industries. If you're interested in credible, objective thought leadership in artificial intelligence, email us at services@emerj.com.

About Iron Mountain

Iron Mountain Incorporated (NYSE: IRM) is the global leader for storage and information management services. Trusted by more than 225,000 organizations around the world, the company helps customers lower cost and risk, comply with regulations, recover from disaster, and enable a more digital way of working.

Iron Mountain InSight is a content services platform that provides actionable business insights and predictive analytics through Machine Learning (ML)-based classification of a company's physical and digital information, which adds structure, context, and meta-data to information to make it more usable. The resulting enriched content can then enable enhanced automated governance and workflow throughout an oil and gas organization.