EXECUTIVE MASTER

STATE OF THE ART REPORT

Artificial Intelligence
The École Polytechnique Executive Master: A UNIQUE COMBINAISON OF TECHNOLOGY AND BUSINESS

The Executive Master is a unique diploma from École Polytechnique targeted to senior managers that combines technology, innovation management and leadership.

This part-time program is designed in modules and allows participants to follow the courses while respecting their professional commitments.

The modules cover various technological and innovation fields such as artificial intelligence, Data Science, biomimicry, IoT, new energies, sustainable development, industry 4.0 or biomedical engineering. The courses are given by researchers working in the laboratories of École Polytechnique and of our partner institutions.

The knowledge acquired throughout the program helps participants to carry out a Team Project and a personal project, the “State of the Art” (SOTA) Report. The objective of the SOTA Report is to produce an updated synthesis on a scientific or technological subject. The document has to be quite short and readable by non-experts.

Over the years, many of these outputs have proven to be both intellectually very stimulating and practically very useful to accelerate the structuration of innovation processes. Hence many readers suggested that we make them available to a larger audience.

This SOTA Report is part of a selection we did and for which the authors – now our Executive Master alumni – gave us their authorization to publish and share.

We hope you will enjoy exploring this “treasure”. Do not hesitate to send us some feedbacks or get in touch directly with the authors.

The École Polytechnique Executive Master team

Executive.master@polytechnique.edu

---

Disclaimer: these reports are written under the sole responsibility of their author.
ARTIFICIAL INTELLIGENCE IN INVESTMENT BANKING
Towards the augmented banker?

SOTA Report, Executive Master, Ecole Polytechnique
September 9th, 2022

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, Investment Banking

Eric ARNOULD
eric.arnould@polytechnique.edu
Executive Summary

Digitization has been a major theme in investment banking (“IB”) for several years. In addition, artificial intelligence (“AI”), machine learning (“ML”), deep learning (“DL”), natural language processing (“NLP”) and many other data processing technologies are increasingly mainstream and widely spread in many industries.

In finance where it is all about data and intelligence, the promise of AI is huge. Some segments like retail banking are already undergoing massive transformation. Still, when it comes to IB, we are only at the beginning of a new age and the best is yet to come.

Where do we stand today? Shall we expect major disruption of the IB value chain and what are the most advanced innovations and technologies to date? Is the “augmented investment banker” poised to be a reality and how, or is it a myth? These are some of the key questions which we expect to sequentially address in this report.

After drawing a picture on the use of AI/ML/NLP in IB over the recent years and the underlying drivers in a first section, we will focus in a second section on state-of-the-art applications and use cases as well as academic research to try and identify the key areas of focus for the coming years. The third part will present a view of what the augmented banker could be, weighing potential benefits and challenges.
1. AI in IB as of today: a powerful yet focused mega-trend

IB experienced rapid digitization over the last decades. This gave birth to a new ecosystem where data is everywhere, from internal info on clients, meetings and transactions saved in internal databases to external stock prices, macro indicators or news and transcripts. This led to the development of many AI use cases over the last decade as AI techniques improved, yet concentrated on some areas and features.

1.1. Powerful drivers to the development of AI in IB

The development of AI in IB is backed by several drivers:

- **Availability of data and adequate quality**
  IB is all about data and data processing. Financial data is mostly quantitative and user-friendly, be it stock prices, macroeconomic, accounting and valuation data points. Textual data is also available in standardized contract documents, and unstructured news articles, thus a useful base for NLP applications.

- **Massive development of so-called “alternative data” and data providers**
  According to Grandviewresearch, alternative data should grow by an impressive 54% p.a. by 2030, at $143bn from $2.7bn in 2021. Financial services are the first users of alternative data, accounting for c.15% of total. Meeting increasing requirements of financial institutions, many providers emerged, from global leaders eg. Bloomberg and Reuters to innovative companies eg. Ravenpack, Preqin, Dataminr or Globaldata etc. They provide corporates and financial institutions with data such as credit card transactions, satellite and weather data, inventories information, web scraped data, social media statistics. IB institutions and their investor clients use this information to detect weak signals and trends, and to back their lending, risk or investment views.

- **Sharp decline in computation cost**
  Thanks to more powerful CPUs, including quantum computing, and to the development of cloud computing and high-speed servers.

- **Perceived benefits, such as time and cost savings, and efficacy**

---

1 This is particularly true for public companies listed on a stock exchange, as opposed to private unlisted companies for which data is usually scarce and hard to find. But this may change in the future on the back of better data scraping methods and disclosure requirements set by the regulators.

2 This is expected to be amplified with the development of quantum computers, some of which are already marketed to financial institutions and hedge funds.
Cost reduction is the #1 reason to support AI deployment in financial services, mentioned by 37% of IB professionals. Time saving is another benefit, including in value-added tasks such as target finding in M&A. According to a study by Accenture, analysts can save up to 60% of their time using AI/NLP to identify M&A targets. Efficacy is another big advantage of AI/ML, for instance in trading, allowing increased accuracy and reduced human mistakes, trades executed at the best possible prices, automated and simultaneous checks.3

- **New regulations such as Europe’s MiFID II directive**
  As an example, the obligation to store data of every transaction for five years, as well as new transparency requirements, mean investors and brokers are obliged to produce various kinds of data and reports for regulators.

- **Raising awareness of AI potential in IB**
  Many consultancies such as Mc Kinsey, Roland Berger, Deloitte or BCG issued papers about the potential of AI in banking. In one of the rare studies focused on IB, LSEG Labs evidences the potential of NLP, applicable to sell-side research analysts and risk/compliance departments, focused on increasing efficiency (automation of research, M&A analysis, predictive analytics) and generating revenue growth (deal origination). Rising awareness of IB practitioners is noted, with 20% considering NLP fundamental and another 20% convinced and investing in this technology.

After a period where retail banking attracted most AI investments to digitize the customer relationship through chatbots, IB is now taking the lead4.

### 1.2. Most frequent AI use cases in IB as of today

Potential to deploy AI solutions exists across all IB activities. IB typically covers Capital Markets (or Sales & Trading), Research (macroeconomic, equity, credit and ESG) and Corporate Finance (M&A, Equity and Debt Capital Markets, Acquisition and LBO Financing) and 2 dimensions from front office (customer-centric) to middle/back office (eg. internal processes and control compliance functions).

---

3 [https://www.mesh-ai.com/blog-posts/where-are-machinelearning-ai-applied-across-investment-banking](https://www.mesh-ai.com/blog-posts/where-are-machinelearning-ai-applied-across-investment-banking)

The next table presents a summary of existing AI applications developed by selected IB houses as of today:

<table>
<thead>
<tr>
<th>IB activities</th>
<th>Capital Markets</th>
<th>Research</th>
<th>Corporate Finance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Client-centric</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithmic trading / BNPP(^5) (ML, NLP)</td>
<td></td>
<td>Geopolitical risk prediction / Deutsche Bank(^6) (ML, NLP)</td>
<td>Equity raising prediction / JPM(^7) (ML)</td>
</tr>
<tr>
<td>Support to traders / Goldman Sachs (ML, DL)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investor profiling / Société Générale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Internal</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fraud detection, anti-money laundering / BNPP, Goldman Sachs, JPMorgan, Société Générale(^8), Crédit Agricole etc. (ML)</td>
<td>Global research engine JPM (NLP)</td>
<td>IPO step guide “Deal link” / GS(^10)</td>
<td></td>
</tr>
<tr>
<td>Trade execution / ING, UBS(^9) (ML)</td>
<td></td>
<td></td>
<td>KYC automation, contract review, automated translation / BNPP (NLP/G)</td>
</tr>
</tbody>
</table>

Lessons may be learnt from the above.

- Most AI use cases to date focus on internal processes rather than client applications, and volume businesses eg. trading rather than tailored activities eg. M&A. This is illustrated in the Refinitiv AI/ML 2020 survey\(^11\), with 66% of respondents mentioning Risk as a key area of focus for AI/ML deployment, vs. 32% for equity or debt raising and 31% for research.

- The level of sophistication rose over time, from basic algorithms to complex ML tools with predictive analytics. Algorithmic trading (including through robo-advisors)\(^12\) belongs to this category, supporting Sales & Trading by recommending baskets of suggested stocks based on performance and liquidity criteria, combined with sentiment analysis derived from news or call transcripts.

- While large houses were the driving force years ago developing proprietary solutions (either through dedicated AI teams, or through developments by professionals

---


7 Emerging Opportunities Engine tool designed to help bankers to find companies expected to need debt or equity raisings in the future, 2017 (https://www.businessinsider.com/jpmorgan-using-machine-learning-in-investment-banking-2017-4?r=US&IR=T)


10 «Deal Link » product advertised by Goldman Sachs as early as 2017 allows to segment an IPO process in 146 steps from KYC to settlement. Once largelye communicated by the US bank as the start of a major change towards IB, automation, the tool is more of an automation attempt than a real AI platform, and more of an internal check list tool than a disruptive customer-centric application.

11 AI/ML Survey, Refinitiv, August 2020

12 https://corporatfinanceinstitute.com/resources/knowledge/other/machine-learning-in-finance/
themselves), AI tools are developed by an increasing number of financial institutions\textsuperscript{13}, and many AI vendors emerged, offering solutions to financial players.

- Corporate finance (M&A, ECM etc.) and Research are still underpenetrated by AI compared to Capital Markets, especially on the front office side. Yet this is poised to change, and recent surveys indicate that sell-side analysts are increasingly receptive to AI.
- NLP is at an early stage in IB, but may be the next big thing, with promising applications in sentiment analysis and beyond, as scraping, tagging and classification of unstructured information will progress (Zimmermann, 2020).

2. State-of-the-art AI draw a new IB paradigm

AI in IB is entering a new phase. This is backed by academic research works. This is also directly supported by practitioners and data vendors themselves, pushing several initiatives either for commercial deployment or internal use. In the following, after a short literature review, we will review a number the latest state-of-the-art research works and practical developments in key IB areas, from Capital Markets to Research (credit or equity), Corporate Finance and ESG – the latter being a field of growing interest in finance.

2.1. AI applications for IB: a literature review

Interestingly, while AI in banking as a whole gave birth to an abundant literature, AI in IB remains far less represented, not to say unexplored. A Google Scholar search for articles with “Machine Learning” and “Investment Banking” returns only 39 articles over the last 5 years, as shown in Chart 1.

\textsuperscript{13} According to a proprietary survey conducted with ODDO BHF to a panel of French and European institutional investors in December 2021, close to 50% of asset managers are said to work on proprietary algorithms of all forms (from simple Excel spreadsheets to advanced algorithms).
Other topics prevalent in IB or Capital Markets such as stock price prediction attracted more articles. The same is true for selected issues such as fraud detection using NLP or AI-powered robo-advisors in retail banking.

Several articles present literature reviews of AI in Banking and similar fields such as Accounting and Audit (Fisher et al., 2016). Among these, the article “Data Analytics in Investment Banks” (Iraqi et al., 2021) offers one of the most comprehensive reviews, including a literature survey, an overview of IB departments involved in AI, opportunities and challenges, and examples of use cases already developed. Some other articles present distinctive angles such as an overview of innovative fintech serving IBs for new use cases (Hunt et al., 2020).

Breaking down the sample of articles by typology and region shows that:

- While most articles are issued by academic researchers, only a handful of articles have been issued by IB professionals themselves. A significant number is signed by data scientists and professionals of large companies or Fintech startups eg. Bloomberg, S&P Global Markets or Alphasense. Of particular interest is the series of articles published since 2017 by S&P Global Markets (Zhao, 2017), covering an introduction and definition to AI and related techniques, basic concepts of data treatment, financial use cases from sentiment to earnings calls analysis and even providing blocks of code to develop simple tools.
- Geography-wise, the production is balanced between the US, Europe and Asia. Some countries like India are more active on topics such as stock price prediction. Analysing specifically the small number of academic articles more relevant to IB topics, the US and Europe are the most active geographies, with all European geographies from Germany to France, Switzerland to Ireland and the Nordics – a signal of hope for Europe in the US-driven IB world.

2.2. Capital markets

Capital Markets drives a lot of research interest on the back of many applications aiming at detecting trading signals through what is generally referred to as “sentiment analysis”\(^\text{14}\). This boasts many use cases from detecting signals from statements by C-level executives to decoding unstructured data such as press releases, earnings calls and social media, to pursuing predictive analytics and enhancing investment returns. This is a key research topic for AI and especially NLP\(^\text{15}\).

**Academic research**

Most frequent fields of recent research cover “Word embedding” (Kamal *et al.*, 2021) and improvement of existing algorithms:

- **Word embedding** is an NLP technique where words or phrases are mapped to real number vectors, using different techniques eg. BoW (Bag of Words), TF-IDF (Term Frequency-Inverse Document Frequency), Word2Vec (a popular method allowing to vectorise words, allowing to do multiple operations like addition, substraction), BERT (Bidirectional Encoder Representation from Transformers, a model published in 2018 by Google based on Transformer, a mechanism learning contextual relations between words), Glove (another unsupervised learning algorithm for vector representation) or FastText (collection for studying word embeddings and content categorization).

- **Algorithm improvement** is another field of research, notably neural networks techniques, from DNN (Deep Neural Networks) to RNN to CNN, LSTM (Long Short Term Memory models) and GAN (Generative Adversarial Networks) (Koshiyama *et al.*, 2021).

\(^{14}\) Sentiment can be defined as the proportion of positive or negative words in a statement, using is often based on standardized classification of words. See for instance the financial dictionary proposed by Loughran and McDonald in “When is a liability not a liability? Textual Analysis, Dictionaries, and 10-Ks”, T. Loughran and B. McDonald, Journal of Finance 66 (2011), 35-65.

\(^{15}\) It is fair to note that as of today, the reality of the positive impact of AI on returns is still to be confirmed. This is an important issue for a wider adoption of AI in Capital Markets.
Capitalising on NLP developments, use cases emerged in Capital Markets addressing portfolio recommendations. While more exploratory, this area is illustrated by two research papers opening new perspectives.

- In the first piece (Deryck et al., 2021), the authors propose an application to assist equity sales with the creation of an investor profile using an NLP interface to analyse the objectives and preferences expressed by the client instead of lengthy questionnaires. This brings several benefits related to time (and cost) savings, reduction of the operational risk linked to automation and to enlarged selection of eligible assets.

- In a second paper (Phalippou, 2021), NLP is used to build an index (listed companies exposed to private equity investment such as KKR, Apollo, Eurazeo etc.) by extracting and selecting companies based on text rather than accounting and financial criteria. An immediate benefit is that the NLP approach is much faster than the extraction of detailed semi-public to private financial data from corporate disclosures. Another difference lies in the definition of the weights of each company in the index, derived from the frequency of mentions in news articles and not from market capitalisation. The last difference, even if the author insists that it was not an objective initially, is outperformance vs. analytically designed indices. This paves the way for new indices which may give birth to new financial products and especially innovative ETF (exchange-traded funds) based on innovative approaches.

**Practitioners’ initiatives – the case of Alphasense**

Financial data vendors are also at the forefront of new developments. A state-of-the-art example of new IB use cases for Capital Markets is Alphasense.

This US fintech founded in 2011 strongly accelerated in recent years and is a leader in data processing applications such as extraction, summarisation and sentiment analysis. As a data vendor, Alphasense offers research tools, like Bloomberg or Eikon, yet enhanced with AI and analytical functions. Alphasense provides earnings calls and transcripts tagged for sentiment analysis, with color codes denoting positive or negative statements. Alphasense also

---

16 Internal research teams at financial data vendors such as Bloomberg are initiating advanced research works focusing on improving NLP techniques. See “4 NLP Papers from Bloomberg Researchers Published in “Findings of the ACL: EMNLP 2021” & Conference Workshop”, 2021. Of interest is for instance the search for designing more clustering algorithms allowing to distinguish genuine relevant documents for a topic versus so-called “topic noise” (“Fanatic: FÅst Noise-Aware Topic Clustering”, A. Silburt, A. Subasic, E. Thompson, C. Dsilva, T. Fares, Bloomberg, 2021).

17 Its success is also evidenced by many awards and massive equity rounds of which a recent $180m Series C in October 2021 attracting Goldman Sachs, Morgan Stanley and Citi.
offers data extraction and summarization with advanced granularity and categorization. As an example, it searches data from different sources categorized according to domains (product information, news, M&A, financials etc.) and then regrouped in summaries created through NLG.

Alphasense uses NLP to identify cognitive signals in investment decisions through a blend of advanced NLP algorithms, augmented intelligence and AI-enabled business insight capabilities. The solution uses “smart synonyms”, a form of semantic similarity matching giving users a best-match search on related terms even if not specifically stated in the query.

The company is at the forefront of NLP and DL research for sentiment analysis, with a focus on parsing the language in transcript documents, revealing essential items and significant changes since previous quarters, down to sentence level. Alphasense uses Transformer-based language modelling architecture to produce representations of language and context. Proprietary Language Model is trained on 8m+ financial and business documents, requiring less human-labeled data for each new content source or industry vertical. The teams collaborate with Carnegie Mellon University to explore new text summarization methods, resulting in a multidomain aspect-based summarization dataset called WikiAsp (Hayashi et al., 2021).

Another promising field of Alphasense research is the tentative to use NLP to “quantify aspects of companies’ strategic vision, and by doing so, build a foundation for strategically modeling the dynamics of strategic innovation in an industry, and its relation to financial results” (Grietzer) using so-called “resonance analysis”.

2.3. Research (Credit or Equity Research)

Cross-asset research is another key area of investigation for academic research and practitioners, with promising IB use cases. It is key to highlight recent works related to using AI to assess credit worthiness and bankruptcy risk and to detect high-potential companies.

---

18 Interestingly, mainly thanks to this function, Alphasense also became a distributor of IB research contributing to modify the traditional value chain of the sector. The company was successful in cooperating with banks and signed partnership with Citi or Goldman Sachs as the first distributor of research to corporates or investors. This highlights another use case of growing importance to IB investor clients, namely that in a context where produced research is abundant, an effective search function is of greater value to research buyers. This tends to disintermediate IB institutions, while raising the value of companies like Alphasense as platform and content providers.

19 A concept derived from works from “Individuals, institutions and innovation in the debates of the French Revolution”, A. Barron, J. Huang, R. Spang and S. DeDeo, 2018, in Proceedings of the National Academy of Sciences of the USA.
**Academic research**

From fundamental articles reviewing DL techniques (Hu *et al.*, 2020) to more specific works, a series of recent research pieces show a growing potential to use non-conventional methods to assess credit risks. As methods develop and researchers insist on their complementarity, the best is yet to come to invent new standards accepted by IBs’ credit committees.

New ways of assessing credit can be imagined with textual information. This is particularly interesting for unlisted companies or for small or medium-sized businesses. In a recent paper (Stevenson *et al.*, 2021), a team of researchers demonstrated the value of text for small business default prediction, showing that text alone extracted from textual assessments provided by a lender and analysed using a BERT model turned into scorings surprisingly effective. Their work even showed that adding traditional data did not improve the additional predictive capability.20

A few models further investigated to anticipate startup success based on quantitative and qualitative data. A key challenge is the source and quality of data for early-stage companies. This led some researchers to focus on data sourced from specialist players eg. CrunchBase. Even with a limited yet public data set, some research works show promising results, achieving unicorn prediction of 90%+ on several methods such as logistic regression, recursive partitioning tree, random forest and gradient boosting.

It is important to put these results into perspective as the success of an innovative startup is by essence unpredictable and relies on human decisions and influence, not to mention serendipity. This led other researchers to propose a dual approach (Dellermann *et al.*, 2017), combining ML processing with collective intelligence and cognitive decision-making process to demonstrate its efficiency under extreme uncertainty. On the analytical side, the authors proposed a taxonomy with 20 criteria split in 5 categories, which can be easily analysed. On the cognitive side, a judgment is requested from two groups of experts and non-experts. Data is then aggregated and evaluated. As the methodology has been made public in their original research work, results are yet to be published in a forthcoming paper.

Prediction of IPO success is another field of recent research. A quantitative research work (Colak *et al.*, 2018) on IPOs tried to establish predictive analytics models to anticipate first-day return

---

20 See also above Earlymetrics for a practical example of algorithm already in commercial use. An obvious key question to be further investigated for the implementation of this credit assessment approach is the nature and availability of data coming from the companies to be measured.
based on 25 factors including offer terms, price revision during the bookbuilding period, number of days between filing and launch\textsuperscript{21}. In addition, other works (Sai Kit Tsui \textit{et al.}, 2021) combine the quantitative analysis with a textual analysis of the IPO prospectus and conclude that the readability of the prospectus is an important criteria of IPO over/under-performance.

Practitioners’ initiatives – the examples of GlobalData and Earlymetrics

Two notable initiatives by private companies can be highlighted here, based on the developments of US GlobalData and France’s Earlymetrics for their unicorn prediction and credit worthiness tools.

US data vendor GlobalData develops and markets a tool called Unicorn Prediction. Based on proprietary quant-based Startup Scorecard tool combined with ML algorithms and predictive intelligence, this app uses 20+ items to calculate a probability to become a unicorn\textsuperscript{22}. Key indicators relate to deal value in each round, average investment per investor, number of top investors per round and company sector. The model has been trained and evaluated using 15 different ML algorithms and eventually a logistic regression algorithm has been chosen. As of December 2021, the model had a 45% accuracy, having been able to predict 210 companies out of 468 companies that became unicorns in 2021\textsuperscript{23}.

Earlymetrics is a French startup founded in 2014. Once a rating agency focused on on-demand ratings for unlisted companies, the company pursued over the last two years new developments towards an AI-based credit worthiness and bankruptcy prediction tool.

Earlymetrics rating methodology is analyzing 30 corporate indicators over 3 pillars: management team, product/project and market/ecosystem. These criteria are assessed by credit analysts during interviews, and are translated into quantitative indicators, then into overall ratings and synthetic company reports. As of today, over 3,000 companies have been analyzed, upon request of companies themselves, or investors asking for an independent opinion in the context of vendor due diligence.

\textsuperscript{21} Even if some of the included factors, such as the arguable “market heat” score, might be further investigated, the study offers an interesting starting point to imagine a quantitative approach as a sanity check to decide on launching an IPO or not.

\textsuperscript{22} A unicorn being defined as a private company with a valuation in excess of $1bn.

\textsuperscript{23} Spotting Future Unicorns: Predicting Startup Success using GlobalData’s AI model, February 4, 2022.
Recently, Earlymetrics also developed new algorithms\textsuperscript{24} to process data to create indicators signaling for growth potential or conversely bankruptcy risk. This product, still under development and to be marketed to banking clients within months, is the result of internal data science conducted in cooperation with French CNAM\textsuperscript{25}, the latter also helping to backtest algorithms and to analyse the accuracy of statistical methods.

The product uses neural networks and unsupervised learning to translate collected data into effective usable indicators such as a probability of default in the next two years. Latest version of the algorithm achieved a 98% accuracy in predicting bankruptcy risk, based on a sample of several thousands of companies.

The product is of interest to IB houses to orient their marketing policy, target new prospects and trigger client discussion on the perceived areas of fragility etc. Another key benefit is expected to come from the increased collection of data as clients will share their internal data with Earlymetrics.

This example confirms multiple trends:
- Investment banks rely on third parties to process data, share internal information, and make judgment calls with direct impact on their client policy and revenue generation\textsuperscript{26}
- AI (here, neural networks) is used to make financial decisions
- An ecosystem approach is key, through the cooperation by Earlymetrics with academic institutions as a lever to boost disruptive innovation and validate scientific approach and be approved by IBs as a valid credit analysis tool\textsuperscript{27}.

\textbf{2.4. Corporate Finance (M&A and ECM)}

Corporate Finance (M&A and ECM) may be the most elitist and sensitive part of IB – truly so, as interpersonal relationships and negotiations are key to make deals happen. In this respect,

\textsuperscript{24} Partnership between CNAM et EarlyMetrics, September 2020, (https://earlymetrics.com/fr/partenariat-rd-le-cnam-et-early-metrics/)

\textsuperscript{25} Conservatoire National des Arts et Métiers

\textsuperscript{26} One of the key limitations is the acceptance by IB houses of the « black box » neural network approach. As the weights affected by the algorithm to any isolated indicator cannot easily be explained or may appear counter-intuitive in some cases, IB clients will have to validate of the accuracy of this approach, and it will require effort by Earlymetrics to justify the relevance of their algorithms on bigger data sets.

\textsuperscript{27} Of interest is also the partnership signed by Earlymetrics in 2021 with Euronext to supply financial information on small and midcap tech companies, available on their website.
Beyond existing initiatives to automate due diligences (Nikolaidis, 2018), M&A seems to be a long shot for AI. Yet many researchers and some companies explore different parts of the business with promising results, helped by the increasing computation capacity and by improving data science techniques.

**Academic research**

Researchers explore the use of ML to scout VC investment opportunities and fine-tune the dealflow funnel (Retterath, 2020). Applying algorithms such as supervised gradient boosted tree to a sample of 80k European early-stage companies not only allowed to design scoring tools to predict the success or failure of these companies in the next 3-5 years, but also to identify some key characteristics of performance, and to narrow the promising part of the deal-flow funnel, generating extra performance compared to a wider portfolio. Noteworthy, the author does not suggest replacing humans with ML-based approaches, but solutions for pre-selection, saving time and resources to concentrate on “the most promising targets and put themselves in the best position to secure these deals”. This is an attractive first step towards an “augmented banker/investor” and a way to scale the VC business, reshaping the role of the M&A banker proposing deal ideas to his clients.

More disruptive, some researchers investigate the ability to develop deal sourcing assistants. An interesting approach is provided by the development of new models using graph-based ML (Venuti, 2021). This adds to more classical tentatives to predict M&A using random forest models or NLP. The basics of this promising method lie on graphs, a “historical mathematical concept core to frameworks like Markov Decision Processes” and defined as “a set of nodes and edges representing relationships between entities”. Capturing not only the information but also the relation between two nodes or entities, this method brings enriched information to traditional vectors and matrices, and allows to position an entity or company in relation to its competitors, partners etc. The work proposes a data acquisition methodology using Wikidata covering all companies of top 30 countries and applies the StellarGraph open-source library of graph-based ML tools. The initial model achieved 82% accuracy on a validation dataset, above previous studies. The author also proposes a roadmap for future work using temporal-based GNNs and

---

28 This study performed in 2018 identifies opportunities for legal due diligence through the automated review of documents using NLP. Financial due diligence is more of a question mark as financial data would be deemed more “stable” and therefore less time-consuming to review. Another question addressed is the heterogeneity of players involved, from legaltech startups to large institutions like JPMorgan with proprietary initiatives.
supplemental NLP extraction, the latter being a promising tool to extract information that a company X may be a direct competitor to company Y – a key data item to assess M&A potential.

**Practitioners’ initiatives – the examples of Refinitiv, Bank of America and Jolt.Ninja**

A key area of AI development in Corporate Finance is matchmaking and target identification\(^29\), be it for M&A deals to find appropriate targets or for ECM to identify most likely investors. Among the most serious initiatives\(^30\), big players like Refinitiv or Bank of America (BoA) or private equity funds engaged significant efforts to build smart AI tools.

Refinitiv developed in 2021 an M&A Prediction model (Aramyan, 2021) to predict M&A targets, combining fundamental analysis (eg. M&A database, stock price history, company fundamentals) with text mining functions to improve overall predictability. After testing different models, Refinitiv opted for **Convolutional Neural Networks (CNN)** and **Bidirectional Encoder Representations from Transformers (BERT)** over **Recurrent Neural Networks (RNN)**, the former performing reasonably well in predicting future M&A targets. Even using these most advanced NLP technologies, success is still moderate, and several issues are still to be tackled, such as the current limitation to listed companies and to English language only.

BoA data scientists created in 2019 a tool called **PRIAM**\(^31\) based on a network of supervised ML algorithms to understand relationships between ECM deals (IPOs, follow-on equity offerings etc.) and investors, and design recommendations on investors best suited to a specific equity offering, thus maximizing demand. 150m data points were collected on 50k ECM deals from public and proprietary data to train the model. Objective is to predict best investors for a deal based on

---

\(^{29}\) Note that others use cases more focused also exist in M&A, such as due diligence automation. This is a use case less focused on «customer-centricity », therefore not developed directly in this report. It is however worth mentioning as a key source of cost and time savings. An interesting example is offered by the «Prepare and Diligence » apps launched by US provider Datasite in 2020, using ML and NLP to simplify and speed up the due diligence process. Product features include: (i) “Smart categorization”, a categorization tool for documents loaded in a deal data room, using a ML engine trained on 3m M&A specific documents, (ii) “AI indexing”, a tool using NLP to suggest to which folder a deal document belong, and (iii) “Intelligent redaction”, an NLP feature to block sensitive information across a whole document based on simple rules. Estimated benefits are critical, as due diligence is estimated to be the #1 M&A success factor by 62% of practitioners, according to a proprietary survey conducted by Datasite in 2020. And 61% anticipate that ML technologies will improve solutions for virtual data rooms.

\(^{30}\) Another original initiative is being conducted by the Shenzhen Stock Exchange to promote matchmaking tools, such as the V-Next platform. The Shenzhen Stock Exchange, known as mainland China’s stock exchange focused on technology companies, launched a cross-border matchmaking platform to connect investors and enterprises, called V-Next. AI is reportedly integrated to optimize the matchmaking functions, pushing investment situations and companies to investors according to their preference. Despite limited roll-out as of today (probably due, to the limited automatization of connecting functions), this platform is an interesting example of innovation in the field of data-driven matchmaking platform in the M&A/ECM space, also showing that AI-driven solutions might affect the role and positioning of the different actors in the value chain – in this example, it is interesting to note that the stock exchange itself, not investment banks, is leading this initiative.

\(^{31}\) PRIAM stands for “Predictive Intelligence Analytics Machine”. Source: Bank of America brings AI to equity capital market, March 2020, Thor Olavsrud, CIO.com.
offering terms, historical deal participation, trading and market data. The tool can score 1,000 investors in seconds and identify most likely investors with 80% accuracy. BoA highlights the combination of long-standing market experience with updated data and analytics as the secret sauce of this product. The project earned a FutureEdge 50 Award. It helps the bank save hundreds of hours previously manually spent by analysts collecting and analysing data.

Starting with modest objectives, other comparable products may be more advanced, such as the Ninja research engine developed by French private equity investment company Jolt Capital. With a view to enhance internal sourcing while saving costs and focusing staff on more value-added tasks, the firm decided as early as 2017 to develop a proprietary sourcing and dealflow tool. Among others, Ninja offers a function allowing to identify and map potential peers (public or private) of a searched company, and to benchmark its competitive positioning vs. peers based on indicators like geographic positioning or intellectual property. The tool is based on NLP using essentially public information scraped from multiple sources and is enhanced through ML functions.

2.5. ESG – a new transversal topic

As ESG emerges as a rising topic in finance and for IB houses, it seems essential to conclude this section with a review of ongoing research and the input of AI solutions. Being a world of qualitative information, ESG brings new requirements, be it for credit assessment or investment recommendation to clients, which may well be the trigger for more radical adoption of AI and NLP techniques in the IB space. Two recent pieces provide complementary perspectives of this trend.

On the credit side, an area of focus for IB houses is the assessment of credit requests from an ESG standpoint, a key criterion in their lending policy as environmental pressure grows. Computing internal ESG ratings is often a must to obtain credit approvals, requiring processing additional and often unstructured information. Recognising that access to financing is essential in

---

32 Interestingly, BoA’s co-head of global capital markets also draws an interesting takeaway resonating with limitations mentioned above in Section 1, i.e. that while “the consumer side of the business tends to be very contemporary because the user base is all on the cloud, (…) on the more corporate-focused side of the house, technology lags, often because clients themselves are on the lagging side of the technology curve.”

33 The tool was already made accessible to the French Direction Générale des Entreprises (a department of the French Ministry of Economy and Finance) and to some teams of Singapore’s sovereign wealth fund Temasek and likely to be of interest to IB professionals and investors.

34 An example close to the author is the “Green Weighting Factor” developed by Natixis as early as 2019 to assign a “colour” to any loan granted by the bank to its clients and established as an internal standard besides more classical
companies taking measures to finance environment-friendly transformations, an international team proposed ML models to anticipate credit ratings of decarbonized firms (Yu et al., 2021). Their finding is that algorithms are relevant to predict credit ratings, and that top influencing variables depend on the credit category (investment grade, crossover etc.) to which these companies belong, opening new perspectives to review internal credit models and support IB clients undergoing climate transition.

Finally, in a paper dated March 2022, a team from Berkeley focused on the increasing need for finance practitioners to have access to consumable ESG information (Mehra et al., 2022). Stating that most of it is in text form spread in multiple reporting documents, they designed a tool for classification tasks of ESG text, using advanced NLP transformer-based BERT model, achieving better accuracy than pre-existent models. Initially focused on Environmental risk scores, their model may also be used to predict Social and Governance scores, covering all ESG dimensions.

Practitioners’ initiatives
Commercial IB applications are yet to be developed in this area and are still in their infancy as of today. Yet several players are investing a lot to develop AI-based products. France’s Ecovadis has developed a sustainability management platform for businesses, powered by AI. Though not directly designed for IB clients, this ML-powered tool product contributes to provide ratings which are of use by many stakeholders, including credit institutions. Valuecometrics is another interesting startup ambitioning to realign values and valuations by enabling investors and companies to better take into account ESG criteria in their activities. The young company developed a tool for investors allowing to get a synthetic view of ESG scores granted by third-party providers as well as high-quality independent metrics to assess extra-financial positioning, through proprietary data scraping and algorithms.

2.6. Interim conclusion: a full backlog of AI in IB apps for the next decade

Abundant research publications and practitioners’ initiatives pave the way for a decade of expanding AI in IB, with different time horizons. The below chart shows a tentative summary to classify projects according to their status, positioned on a typical Gartner hype cycle curve. While some ongoing use cases are still at an early stage, and current available apps are more process-financial metrics. See press release dated September 23, 2019: “Natixis rolls out its Green Weighting Factor and becomes the first bank to actively manage its balance sheet’s climate impact.”
oriented, a full pipeline of customer-centric IB applications is getting closer to commercial roll-out. In other words: prepare for the change!

Chart 2 – Synthesis of AI in IB

3. Towards the augmented banker – myth or reality, and roadmap

As seen previously, AI has the potential to deeply reshape IB. As of today, some parts of the industry such as internal processes for anti-money laundering, compliance and soon credit approval are undergoing more powerful changes than ever before. A lot is still to be done, and IB essentially remains, at least in some businesses like M&A, an activity driven by human and interpersonal skills. As illustrated in latest academic research papers, a hybrid approach combining the best of AI with human, cognitive skills appear to be the winning mix.

3.1. AI in IB: an irreversible trend with tangible benefits

Over the last couple of years, significant progress was made in nearly all fields of IB, touching nearly all departments of IB. Fintechs grow like never before, led by new players with full freedom to explore new fields and zero risk of cannibalising their existing business. IB houses themselves face tougher financial equations, exacerbated by strict regulation driving higher cost of doing business, and pressuring to explore cost savings opportunities.
Against this backdrop, moving towards more AI in IB is not a question of if but when. The next decade, capitalizing on the quantum leap in computation power and AI developments, is likely to see drastic changes in a sector largely stable since the invention of modern equity raising and derivatives trading in the 1980s. Evidences can already be seen in the overnight digitization of IPO roadshows during the Covid-19 pandemics or with the emergence of new players such as PrimaryBid, an innovative UK startup developing an app to facilitate individual investors subscription in accelerated equity placement – a premiere in the oligopolistic equity underwriting business.

AI in IB also brings several tangible benefits, such as:

- **Fluidify corporate financing, both debt and equity**
  Debt-wise, ML already brings improved credit assessment tools. The new frontier will be to facilitate financing of small- and medium-sized, unlisted, early-stage companies with the design and use of new credit scoring methods exploiting unstructured data. This will require banks to approve new internal risk models and procedures, which may take some time. Competition between IB houses to capture business with these young unicorns is likely to help accelerate this process.
  AI-powered databases may give investors new information tools to support investment decisions, educate ahead of IPOs, and complement or replace traditional sell-side equity research provided by IBs – a model already challenged in the context of European MiFID2 regulation. This will enlarge the audience of addressable potential investors – to the benefit of companies seeking for financing – beyond the pool of mainstream big investors, irrespective of their size and existing relation with dominant IBs.

- **Increase the probability of success of M&A transactions**
  The two biggest risks in M&A are execution and lack of fit. AI brings a decisive contribution to solve both, in automating/accelerating the due diligence process, and in contributing to form views, through NLP analysis of unstructured data, of the theoretical match between the acquirer and the target, their management teams, client base etc.

- **Accelerate the ESG transition**
  A new mainstream trend as climate concerns gain momentum, ESG becomes a key topic in every IB business, from financing to equity story to capital raising. Due to its largely qualitative, unstructured nature, ESG data is the ideal candidate for NLP, opening opportunities for new innovative scoring applications useful to IB and their clients.

3.2. Challenges remain – a new ecosystem is required
Having in mind the disruption implied by the 4th industrial revolution (Schwab, 2016), and the potential to automate jobs, one could wonder why the AI transition of IB is not more advanced.

Beyond the importance of interpersonal skills in several IB businesses, several factors may be mentioned, such as structural industry organisation, led by a few powerful players from the US, Europe and Asia with established positions and superior profitability, allowing to absorb external shocks to the benefit of the stability of the banking system.

Yet leading IB houses (universal banks or pure players) are also increasingly aware that AI can be a source of profit and competitive advantage and have started to invest massively. This is a first step. Many others are required to build an AI-friendly IB ecosystem, among which:

- **Further enhance computation power to process massive amounts of data**
  
  This may come in the next years with the commercial deployment of quantum computers, some of which are already gaining traction, with uses concentrated in financial trading at this stage, and possibly extended to other ML/NLP processing applications soon.

- **Set governance standards for the use of AI in banking**
  
  One of the current criticisms formulated against AI-powered credit scoring is the lack of explainability of ML models, notably neural networks. This might even be incompatible with existing regulation where algorithms must be fully understood throughout their lifecycle, as per IOSCO or other regulations such as GDPR.

- **Harmonise and digitize financial reporting with the support of the regulator**
  
  As unfriendly as they can be, US Edgar filings have been a major source for NLP deployment. New regulations are expected to be introduced in Europe in the coming years, paving the way for more standardisation, digitisation and categorisation of financial reports for a large number of companies, mainly listed but also extended to a number of unlisted companies. This will help develop new tools to process information.

- **Disclose more information on growth companies**

  Much information is already available on most promising startups and scale-ups, yet unevenly distributed to investors and the public. As an example, companies included in

---

35 “The fourth industrial revolution”, Klaus Schwab, World Economic forum, 2016. See in particular Tables 1 and 2 highlighting the key transformations by 2025 and examples of jobs most exposed to automation.

36 In an article on potential social consequences of the automation of the banking industry, it is estimated that in the US alone, 1.2m jobs may be at risk in banking by 2030, and 460k in investment management, of which 96k financial managers, 13k compliance officers and 250k loan officers (including credit analysts). Though this estimate overlaps retail banking and is not focused on IB, it shows that the disruption potential is huge and raises questions on the role of more executive jobs. “How artificial intelligence is reshaping jobs in banking”, Penny Crosman in The American Banker, 2018.

the French Tech’s Next40 and FT120 indices shared financial and non-financial information with the French Tech association as well as with Bpifrance and some other institutions as part of their application process. Giving more publicity to even a fraction of this information would provide a useful dataset to feed AI algorithms trying to determine credit worthiness and unicorn potential and would also contribute to better investor education.

Conclusion – potential, limitations and perspectives

AI has the potential to reshape IB in the next decade. After transforming internal process with tangible cost and efficiency benefits, after the ongoing changes in credit assessment, the next frontier is to transform the way front-office bankers interface with their clients, in activities less affected to date such as Debt and Equity Capital Markets or M&A.

AI is not and will not be the alpha and omega of IB innovation. A hybrid model combining advanced AI-powered tools (especially NLP as the processing of unstructured data becomes more accessible) with collective intelligence and interpersonal skills, is the most likely and desirable model at the service of augmented bankers. This would be for the better of the IB industry and their clients, from blue chips to innovative ventures. This may bring fluidity in equity and debt financing, enlarged investor audience and better selectivity in M&A transactions.

With this will also come a new IB landscape with new players, and this may a chance for Europe. US players have always been well positioned in this industry and will continue to drive innovation, but many European companies and initiatives are offering innovative services for the first time in years at several stages of the IB value chain, from debt/equity intermediation to research. They operate in rejuvenated favorable ecosystems driven by smart political initiatives such as development of the French Tech or the Tibi initiative in France, or the Scale Up Europe program at the European level.

Better information, connection and transactions through smart AI applications at the service of corporates and investors – this sounds like a great opportunity for European investment banks!

(39,689 characters)

---

38 This is a topic for discussion to be further investigated in France for example with entities such as Paris Europlace and representatives of public institutions such as the Direction Générale du Trésor (department of the French Ministry of Economy and Finance) or the French Tech business club.
Bibliography

Books & Reports

Deloitte Insights (2021). Making the investment decision process more naturally intelligent / How AI technologies are improving man-machine communication with natural language processing, report.
Forbes.com (2020). The state of AI adoption in financial services.
Refinitiv (December 2021). NLP in Financial Services, LSEG Labs, Refinitiv.
Roland Berger (2019). AI in Financial Services, how to tackle the big buzz, report.
The Economist Intelligence Unit (2020), The Road Ahead, Artificial Intelligence and the Future of Financial Services, report.

Articles

Dellermann, Lipusch, Ebel, Popp, Leimeister (2017). Finding the Unicorn: Predicting Early Stage Startup Success through a Hybrid Intelligence Method, University of Kassel / University of St. Gallen (Germany/Switzerland).
Deryck M., Comenda N., Coppens B., Vennekens J. (2021), Combining Logic and Natural Language Processing to Support Investment Management”, KU Leuven (Dept. Computer Science and Institute for AI), Coppens and Partners Consulting.


Grietzer P. Decoding Sentiment, NLP and Innovation, Alphasense.


Mehra, Louka and Zhang University of California, Berkeley (2022). ESGBERT: Language model to help with classification tasks related to companies environmental, social and governance practices.


Stevenson M., Mues C. and Bravo C. (2021). The value of text for small business default prediction: a Deep Learning approach, University of Southampton (UK), University of Western Ontario (Canada).


Zimmermann V. (2020). How the large investment firms are using NLP.

(5,560 characters, spaces included)
Glossary

Please note that several definitions below are quoted from Wikipedia. After a careful review of alternative definitions of AI and related concepts, it was deemed relevant to use Wikipedia a good summary and synthesis of multiple sources and definitions. The author directly contributed to edit several Wikipedia articles related to corporate finance items such as investment banking and equity capital markets.

Artificial intelligence (AI)

Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by animals including humans. AI research has been defined as the field of study of intelligent agents, which refers to any system that perceives its environment and takes actions that maximize its chance of achieving its goals.\(^39\)

The term "artificial intelligence" had previously been used to describe machines that mimic and display "human" cognitive skills that are associated with the human mind, such as "learning" and "problem-solving". This definition has since been rejected by major AI researchers who now describe AI in terms of rationality and acting rationally, which does not limit how intelligence can be articulated.

Machine learning (ML)

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.\(^40\) It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of

\(^39\) Definition of AI as the study of intelligent agents, drawn from leading AI textbooks. Poole, Mackworth & Goebel (1998, p. 1), which provides the version that is used in this article. These authors use the term "computational intelligence" as a synonym for artificial intelligence.

machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

Deep learning

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised.41 Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.

Natural language processing (NLP)

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Natural language generation (NLG)

Natural language generation (NLG) is a software process that produces natural language output. While it is widely agreed that the output of any NLG process is text, there is some disagreement on whether the inputs of an NLG system need to be non-linguistic. Common applications of NLG methods include the production of various reports, for example weather and patient reports; image captions; and chatbots.

**Supervised learning**

Supervised learning (SL) is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.\(^\text{42}\) It infers a function from labeled training data consisting of a set of training examples.

In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way (see inductive bias). This statistical quality of an algorithm is measured through the so-called generalization error.

**Unsupervised learning**

Unsupervised learning is a type of algorithm that learns patterns from untagged data. The hope is that through mimicry, which is an important mode of learning in people, the machine is forced to build a compact internal representation of its world and then generate imaginative content from it. In contrast to supervised learning where data is tagged by an expert, e.g. as a "ball" or "fish", unsupervised methods exhibit self-organization that captures patterns as probability densities or a combination of neural feature preferences.

**Neural networks**

Artificial neural networks (ANN), usually simply called neural networks, are computing systems inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives signals then processes them and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different

---

transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.

**Convolutional neural networks**

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of artificial neural network (ANN), most commonly applied to analyze visual imagery.43

**Recurrent neural networks**

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed or undirected graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

The term "recurrent neural network" is used to refer to the class of networks with an infinite impulse response, whereas "convolutional neural network" refers to the class of finite impulse response.

**Transformers**

A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of natural language processing (NLP) and computer vision (CV).

Like recurrent neural networks (RNNs), transformers are designed to process sequential input data, such as natural language, with applications towards tasks such as translation and text summarization. However, unlike RNNs, transformers process the entire input all at once. The attention mechanism provides context for any position in the input sequence. For example, if the input data is a natural language sentence, the transformer does not have to process one word at a time. This allows for more parallelization than RNNs and therefore reduces training times.

Transformers were introduced in 2017 by a team at Google Brain and are increasingly the model of choice for NLP problems, replacing RNN models such as long short-term memory (LSTM). The additional training parallelization allows training on larger datasets than was once possible.

---

This led to the development of pretrained systems such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), which were trained with large language datasets, such as the Wikipedia Corpus and Common Crawl, and can be fine-tuned for specific tasks.

**BERT (Bidirectional Encoder Representations from Transformers)**

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google.

BERT is at its core a transformer language model with a variable number of encoder layers and self-attention heads. The architecture is "almost identical" to the original transformer implementation in Vaswani et al. (2017).

When BERT was published, it achieved state-of-the-art performance on a number of natural language understanding tasks. The reasons for BERT’s state-of-the-art performance on these natural language understanding tasks are not yet well understood. Current research has focused on investigating the relationship behind BERT’s output as a result of carefully chosen input sequences, analysis of internal vector representations through probing classifiers, and the relationships represented by attention weights.

**Sentiment analysis**

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. With the rise of deep language models, also more difficult data domains can be analyzed, e.g., news texts where authors typically express their opinion/sentiment less explicitly.

**Investment banking**

Investment banking denotes certain activities of a financial services company or a corporate division that consist in advisory-based financial transactions on behalf of individuals, corporations, and governments.
Traditionally associated with corporate finance, such a bank might assist in raising financial
capital by underwriting or acting as the client's agent in the issuance of debt or equity securities.
An investment bank may also assist companies involved in mergers and acquisitions (M&A) and
provide ancillary services such as market making, trading of derivatives and equity securities,
FICC services (fixed income instruments, currencies, and commodities) or research
(macroeconomic, credit or equity research).

Venture capital
Venture capital (VC) is a form of private equity financing that is provided by venture capital firms
or funds to startups, early-stage, and emerging companies that have been deemed to have high
growth potential or which have demonstrated high growth (in terms of number of employees,
annual revenue, scale of operations, etc). Venture capital firms or funds invest in these early-stage
companies in exchange for equity, or an ownership stake.

Mergers & acquisitions
In corporate finance, mergers and acquisitions (M&A) is an investment activity consisting in
advising clients (corporates or financial institutions) in transactions in which the ownership of
companies, other business organizations, or their operating units are transferred or consolidated
with other entities. M&A starts by analysing and suggesting to clients potential mergers with, and
acquisitions of, other enterprises with a view to grow or downsize, and change the nature of their
business or competitive position. M&A activity also covers the execution of these transactions in
all related financial, accounting, due diligence or legal aspects from initial contact to settlement
of the transaction.

Equity Capital Markets
In corporate finance, Equity Capital Market is an investment banking activity consisting in
advising companies, also referred to as issuers, to raise equity on capital markets. ECM consists
in preparing the equity issues, from designing the equity story and marketing materials of the
proposed transaction to placing the underlying equity securities to institutional and retail investors
through an adequate marketing strategy. Equity securities placed by an ECM desk can range from
common shares to convertible bonds into shares, the latter sometimes designed as Equity-Linked
Capital Market.

Alternative data
Alternative data (in finance) refers to data used to obtain insight into the investment process. These data sets are often used by hedge fund managers and other institutional investment professionals within an investment company. Alternative data sets are information about a particular company that is published by sources outside of the company, which can provide unique and timely insights into investment opportunities.

Alternative data sets are often categorized as big data, which means that they may be very large and complex and often cannot be handled by software traditionally used for storing or handling data, such as Microsoft Excel. An alternative data set can be compiled from various sources such as financial transactions, sensors, mobile devices, satellites, public records, and the internet. Alternative data can be compared with data that is traditionally used by investment companies such as investor presentations, SEC filings, and press releases. These examples of “traditional data” are produced directly by the company itself.

Since alternative data sets originate as a product of a company's operations, these data sets are often less readily accessible and less structured than traditional sources of data. Alternative data is also known as “exhaust data.” The company that produces alternative data generally overlooks the value of the data to institutional investors. During the last decade, many data brokers, aggregators, and other intermediaries began specializing in providing alternative data to investors and analysts.

**MiFID Directive**

Directive 2014/65/EU, commonly known as MiFID 2 (Markets in financial instruments directive 2), is a legal act of the European Union. Together with Regulation (EU) No 600/2014 it provides a legal framework for securities markets, investment intermediaries, and trading venues. The directive provides harmonised regulation for investment services of the member states of the European Economic Area — the EU member states plus Iceland, Norway, and Liechtenstein; the United Kingdom will continue to implement the directive during the transition period. Its main objectives are to increase competition and investor protection, and level the playing field for market participants in investment services.