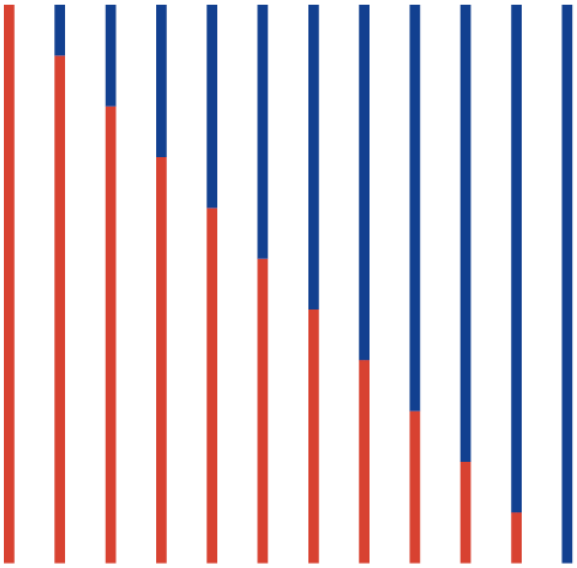


JANUARY 2023

AUTOMATED ANALYSIS OF RISK FACTORS PUBLISHED BY LISTED COMPANIES: A USE CASE OF NATURAL LANGUAGE PROCESSING FOR THE AMF

CORENTIN MASSON

SUMMARY

The entry into force of the "Prospectus" Regulation in 2019 increased issuers' obligations of disclosure concerning their risk exposure when publishing their Universal Registration Document (URD). Since it is incumbent on the AMF to supervise proper compliance with the requirements regarding the presentation of risks, the regulator wanted to explore the potential offered by Natural Language Processing (NLP, a domain of Artificial Intelligence) technologies for analysing these documents more efficiently.

The AMF's work concerned the application of recent breakthroughs in deep learning, which could, in particular, manage the semantic complexities of the risk factor sections, such as interlinking between certain events (e.g., a geopolitical risk could result in a risk of a surge in energy prices") and their specific forms in each sector of activity. Moreover, the use of these techniques would also enable the AMF to facilitate comparison of the risks of issuers of various nationalities by managing a multilingual set of documents.

Based on more than a hundred financial institutions URDs between accounting years 2012 to 2020, the experiments show, in particular, that it is possible to automatically assess the breakdown of risks by sector or by issuer, to monitor their changes over time. It is also possible to detect the most significant variations from year to year in the level of references to each of the risks described. As an illustration, the tool developed visually highlights the emergence of descriptions of the pandemic risk in the URDs published in 2021 (for accounting year 2020), and the constant growth in IT security risk or temporary increases in regulatory risks in certain sectors (e.g. due to the Benchmark Regulation or the 2013 Act on the separation of banking activities).

At the end of this initial experimental phase, but without understating the difficulties that would be involved in generalising this tool for all issuers in the French market, the first results would seem promising and allow the AMF to envisage eventually using NLP techniques more extensively to support its work: not only in the monitoring of the information reported to the market by issuers (and in assessing the quality of this reporting), but also to contribute to the production of thematic research.

Finally, to facilitate the automatic processing of issuers' documents or of regulatory documents more generally, the AMF wants to promote the use of machine-readable formats and, especially, to combine this with good practices to optimise its use.

1. THE IMPLICATIONS OF AUTOMATED EXTRACTION OF RISK FACTORS FOR THE AMF

The AMF, in its role as regulator of listed companies (alternatively referred to hereinafter as "issuers" of listed securities), monitors the quality of the financial and non-financial information published by these companies in accordance with their (periodic and permanent) obligations, and at the time of their financial operations.

Financial and non-financial information is an important factor in decision making by investors in the markets. Among the numerous factors for which issuers have an obligation of disclosure is their exposure to risks assessed as important. This information must be precise and transparent in order to contribute to the investment decision.

Since 2019, the European regulations have laid down, in particular, new formal requirements for the presentation of risk factors, by constraining both the form and the order of presentation and by adding, *inter alia*, criteria of materiality and specificity[1] (the "Prospectus" Regulation (EU) 2017/1129 repealing Directive (EU) 2003/71/EC). For the application of this new Regulation, the AMF is tasked with checking the compliance of the "Risk Factors" sections.

This examination is currently entirely manual, and can prove to be long and tedious (several hundred documents that must be analysed to cover all issuers on the French market). Therefore, the development of an automated tool for risk factor analysis, including both artificial intelligence techniques to extract information from the text and a set of appropriate displays of the extraction results, could usefully support the work carried out by the AMF as part of its duties.

Three main interests were identified:

- the initial experiments showed that it was possible to extract the most relevant sections for each risk factor described by an issuer, and then present them to the AMF staff on a single screen for consultation (example of the low-interest-rate risk of an issuer in Annex 1). By then proposing to consolidate these sections in a single display for all the issuers of a given sector, for example, the staff would be able to **save time on an initial phase of qualitative analysis** of the description of risks for regulated entities supervision.
- the tests conducted on the artificial intelligence models applied to processing the text of the "Risk Factors" sections also showed the possibility of **systematising thematic analyses, and extending them to all issuers** (see below, sub-section 1.3 Results), particularly in order to:
  - compare against its sector how an issuer reports on its risk factors (comparative sector study, *see Figure 3 and Annex 2*);
  - conduct thematic analyses on a given type of risk and analyse how it is presented for a set of issuers (e.g. climate risk factors in the financial sector, detailed study by risk. *(See Figures 3 and 4 for examples of relevant charts for such a study*).
  - assess the change over time in the risk factors of one or more issuers thanks to the easy visualisation of trends (*see Figure 5 below*), making it possible to identify emerging risk factors from one accounting year to the next (such as, for example, the risk related to the health crisis, *see Figure 4 below*);
- future investigations on the automated analysis of risk factors could make it possible to conduct consistency studies regarding the main items stressed in the issuers' various communication channels and strengthen the activity of monitoring the quality of the information communicated to the market: for example by cross-checking issuers' information regarding their risk factors in their regulatory documents with the information transmitted via the other financial and non-financial communication channels (particularly press releases).

[1] These criteria are defined by Article 16 of the Prospectus Regulation and are clarified in the ESMA guidelines concerning the Risk Factors sections, available at this address.

Focus on academic research in finance using NLP

Although the dissemination of information in the financial sphere has been theorised and studied for many decades, automated processing (including reading and analysis) of documents has been more belated. Since the first experiments in the early 2000s, we have identified three main lines of research, which are distinguished mainly by the questions that they try to answer.

The first line includes work with the main objective of processing the data extracted automatically from text documents (newspapers, earnings announcements, regulatory documents, etc.) in order to examine how the information communicated is disseminated in the financial sphere, alters investors' perceptions or can inform us about the information included in prices. The aim is to obtain a better understanding of the relationship between the real economy and the financial sphere by examining the content of the information available in natural language. For example, by quantifying the risk premiums of exposure to climate change automatically, researchers show that these premiums are due more to opportunity shocks than to physical or regulatory risks; others note that giving specific details of risks in the regulatory documents reduces the magnitude of future shocks. The second line is closely related to the first, but focuses on the inclusion of information extracted from the text data in portfolio optimisation, volatility prediction or asset pricing models. For example, Engle et al. (2019) described an approach allowing for protection against climate risk based on news. J. Lu and X Huang (2021), for their part, proposed using event detection to predict the price of crude oil. The last line of research endeavours to analyse the quality of the text content by examining the regulatory documents in particular. This includes, *inter alia*, work on the readability of the texts, their tone and other techniques to make the content more favourable to the issuer who presents it.

Note that the text analysis of risk factors is especially profuse with regard to the SEC's 10-K filings for which there is a very substantial mass of data readily available[2] in a format very conducive to machine processing.[3] However, since the start of 2022 the French URDs now published in XHTML format are far less conducive to processing except for the sections subject to the European Single Electronic Format (ESEF)[4] (cf. "Focus on the implications of an optimal use of machine-readable formats").

Lastly, numerous academic publications[5] are available regarding the influence of the description of risks on the stock market returns for an issuer, on how these risks are perceived and processed by investors, or again regarding the performance gains that can be provided by this data to protect a portfolio from certain identified risks.

[2] All the documents are available via the EDGAR platform on which it is possible to do a manual or automatic search via an API (Application Programming Interface).

[3] Inline XBRL, XHTML with a standardisation constraint applying to the whole document.

[4] Cf. description of the implications of the ESEF on the ESMA website.

[5] Such as “Overlapping Narrative Risk Disclosures and Return”, J. Bai et al; “Revealing the Risk Perception of Investors using Machine Learning”, M. Koelbl et al.; “Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns”, A. Lopez-Lira.

Focus on the implications of an optimal use of machine-readable formats

In the "Open Data" Directive (EU) 2019/1024,[6] the European Commission defines the document formats that allow automatic processing by a machine (machine-readable formats), and promotes their use in order to facilitate access to data.[7] Of these formats, "XHTML", for example, allows a document to be written using a system of markers (or "tags") to define both the structure of the content (title, section, sub-section, etc.) and the referencing of certain specific information within the text (or a table).

While the regulations aim above all to define the key information that should be tagged (particularly within the framework of the ESEF), in practice, when the XHTML format is used, the use of markers concerning the document structure more particularly is crucial to allow processing of the document as a whole. For example, the markers defining sections or title levels allow a machine to navigate more easily in a document of several hundred pages in order to isolate a portion of the document, such as a section in particular. Similarly, the XHTML format provides for table structure markers which are used to facilitate data retrieval without having to use sophisticated techniques such as image recognition.

Accordingly, the use of the XHTML format alone is not sufficient for documents to be able in practice to be machine-processed entirely, i.e. not just the marked-up key information. Thus, the attempts to use 2022 URDs published in XHTML format for the analysis of risk factors (which are not key information marked up pursuant to the regulations) were not conclusive notably due to the absence of markers for the document structure or the construction of tables and the use of markers for the placing of words in the document.[8] Because of these inappropriate processing operations, identifying a particular section of the document and processing tables is very complex. Even worse, words and paragraphs may appear in disorder compared with the visual rendering and words may be cut or merged with others, thereby making the content of certain documents almost unusable.[9]

The AMF would like to draw the attention of document producers to the importance of complying with good practices to optimise the quality of XHTML files, and in particular: the use of appropriate markers to separate sections and paragraphs and to structure tables, and the banning of practices which make it possible to invert the order of sentences or words in the code by comparison with the visual rendering.

2. EXPERIMENTS: APPROACHES AND RESULTS

2.1. PRESENTATION OF RISKS BY ISSUERS

In the section of URDs dedicated to "Risk Factors", issuers describe the various risks to which they are exposed as a series of random events that could have a significant negative impact on their earnings or their growth.

There are numerous types of risks, such as climate risks and regulatory risks, which can themselves be subdivided into several separate risks: in its 2020 report, for example, an insurer described the risk of regulatory change under five aspects, and in particular: "Capital requirements", "Issues related to money laundering and corruption", "Benchmark reform" and "Changes in the IFRS standards". In practice, each of these variants constitutes a risk in its own right.

Moreover, issuers do not merely list the risk factors separately one after another, but also stress the very strong interlinking of the risks with one another. For example, an interest-rate hike could lead to a credit risk; or a geopolitical risk could lead to a risk of a surge in energy prices (see below, "Focus on a risk paragraph in the 2020 URD of an insurer" for an example of a risk paragraph incorporating interlinked aspects of several risks).

---

[6] Article 2, paragraph 13.

[7] The promotion of this type of format is also found in EFRAG's work on *european sustainability reporting standards* (ESRS), and in particular in its proposed general requirements: see *DRAFT ESRS 1 General Requirements* (link).

[8] Which results in numerous errors in the extraction of text, such as inversion of the order of paragraphs, words and/or letters, the concatenation or splitting of certain words and the disappearance of letters.

[9] These defects are due to the conversion of Word documents to XHTML format by third-party software programs, which do not use markers for their roles stipulated by the HTML and XHTML standards of the *World Wide Web Consortium* (link).

Finally, although the ESMA guidelines specify how the regulatory requirements relating to risk factors should be implemented, in practice it can be observed that the risks are explained in various ways, notably:

- either to take sector specificities into account (e.g., operational or competition risks can prove to be of very different kinds depending on whether the issuer is a bank, an insurer or a public works company);
- or in their levels of detail (in its URD concerning the 2020 financial year a bank described a credit risk relating to its lending activities and a credit risk relating to its bond holdings, while an insurer distinguished between two credit risks, the first concerning corporate bonds and the second concerning sovereign bonds).

So, while an analyst would be able to interpret the numerous items of information in this section without too much difficulty, automating its analysis by a machine represents a real challenge.

Focus on a risk paragraph in the 2020 URD of an insurer.

"The Group's results could be significantly affected by the economic and financial situations in Europe and other countries around the world. The threat of a global economic depression due to public health, cyclical and/or commercial reasons (e.g., the ongoing US-China trade war) remains, and a lasting macroeconomic deterioration could affect the company's activities and results. The current low interest rate environment is reaching previously unknown levels and, in the event that interest rates rise, the current exceptional level of indebtedness would become a source of major financial instability. Current monetary policy seems to have reached a point where any additional easing would probably have little significant economic effect. These trends could result in financial markets experiencing a period of very high volatility, with consequences including waves of corporate bankruptcies and potentially sovereign defaults in vulnerable regions, a fall in the value of the main asset classes (bonds, equity, real estate), and even a major liquidity crisis. In the absence of a quick and mass roll-out of vaccines against Covid-19 to the general population, the economic outlook remains negative. In addition, the current decline in the US economy and continuing economic disparities between European countries might have further political and economic impacts. For further information on investments, see Section 1.3.9.2 – Net investment income and investment income on invested assets, and Section 4.6 – Notes to the consolidated financial statements, Note 8 – Insurance business investments."

This paragraph contained in the sub-section "Risks related to the macroeconomic environment", and in particular the portion on "Risks of a deterioration of financial markets and the global economy", is a typical example of the interlinking of risks. Here the issuer depicts qualitatively the tensions weighing on the economic environment: risks concerning the economic situation (blue), rising interest rates (yellow), default (orange), continuation of the Covid-19 pandemic (green) and liquidity (purple). Each of these factors could spiral into a price risk (red).

2.2. THE APPROACH ADOPTED FOR AUTOMATING RISK ANALYSIS BY A MACHINE

In order to help the AMF in its work of risk factor analysis, the tool must first be able to:

- identify the main risks present in each of the documents published by the issuers;
- estimate the materiality of each of these risks; and
- reference the paragraphs of the document relating to each risk.

The URDs are very dense documents (several hundred pages) containing various sections including that on risk factors (more than about ten pages on average). The machine must therefore be capable of identifying the precise portions of the document discussing risks. For this purpose, an initial technical brick[10] has been developed in order to identify the relevant paragraphs and pages.[11]

Then, among the paragraphs retained following this first stage, the tool must distinguish between the various types of risks mentioned there. To do this, each risk can be seen as "a theme", characterised by the presence of a set of words belonging to the same lexical field. "Theme detection" is a common task of NLP work, and numerous models[12] based on research on lexical fields have already been proposed in the academic literature.

[10] Corentin Masson and Syrielle Montariol. 2020. Detecting Omissions of Risk Factors in Company Annual Reports. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 15–21, Kyoto, Japan.

[11] In 2022, work is also under way to reconstruct the contents of each document, based on the FinTOC competition proposed within the framework of the *Financial Narrative Processing Workshop* co-located at LREC 2022 (link).

[12] See Annex 3 for more details on the "Theme Detection" models.

On the basis of the existing models, a specific algorithm has been developed for risk factor analysis in which each theme is then associated with a risk (see Table 1 below).

Table 1: Examples of identified risks

Lexical field	**Associated risk**
attempt – IT - intrusion - confidential - cyber - attack - malicious - hacking - obsolescence - cyberattack	"Cybercrime risk"
transition – investment* - footprint – change – coal - climate-related - environmental - hydrocarbon – carbon - esg	"Climate risk"
contagion – uncertainty – global – measure – natural – appearance – transmission – virus – coronavirus - wave	"Pandemic risk"
rate - variation – investment* - currency - fluctuation - duration - exchange rate - value - bond - yield	"Interest-rate/ exchange-rate risk"
fine - law - dispute - diverging - disclaimer – annual* - applicable - constant - penalty - corrective - code - text - scope - adoption - or even*	"Non-compliance risk"

NB: The words marked with an asterisk "" in the above table are not specifically related to the lexical field of the risk to which they are linked. This is because the model used is based on statistics of joint occurrences of words in order to identify the themes, some of which are not always relevant (e.g. "or even" and "annual" which appear frequently in paragraphs relating to "Non-compliance risk", and "investment" which appears in both the "Climate risks" and "Interest-rate risks" paragraphs but not in the others).*

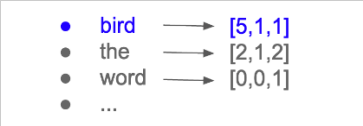
In practice, for each paragraph analysed, the algorithm will return the themes detected (in other words the risks; see Table 1 above) with a confidence probability. Given that the study is at present restricted to documents in French, and following a detailed performance assessment, the "theme detection" model that was finally implemented is a flexible, state-of-the-art model particularly suitable for our requirements. This model is called SCHOLAR and was configured, trained and validated on our data.

Lastly, the machine must analyse the information obtained concerning the paragraphs to estimate what significance is assigned to each risk (i.e. theme). In practice, this means, in particular, being able to associate two (or more) information items with one another, without them being necessarily close to one another in the text. In order to obtain a breakdown of the weight of the risks mentioned by an issuer, the tool will compare the size of the paragraphs devoted to the description of each one. In other words, it considers that the significance of a risk is directly proportional to the quantity of text devoted to it.

Focus: "How does the machine process the text?"

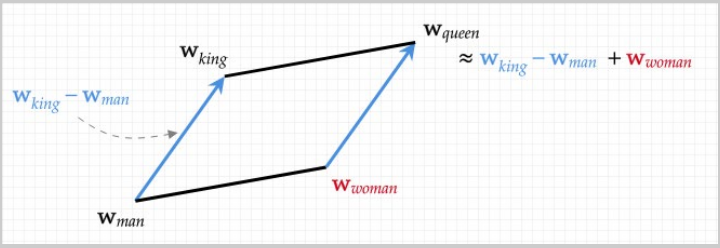
As a string of characters, words and sentences, a text is not directly conducive to machine processing: a preliminary numerical conversion is necessary, commonly called "vectorisation" (see Figure 1 below).

Figure 1: Vectorisation or numerical conversion of words



There are various approaches to performing this numerical conversion, and sophisticated approaches can preserve several properties of meaning and syntax. For example, it has been demonstrated[13] that some of these numerical vectors identify analogies by linear relations: the words *king*, *woman* and *man* can be used to calculate the vector representing the word *queen*[14] (see Figure 2 below).

Figure 2: Linear analogy between word vectors



Three vectorisation approaches were studied for the work on risk factor analysis:

- "bags of words":[15] in its simplest version, the converted text is then represented by the histogram of occurrences of the words forming it;

- "non-contextual word embedding" based on the "skip-gram" algorithm: each word is represented by a vector and numerical conversion should make it possible to preserve the information that two words are similar in meaning;

- "contextual word embedding": a multilingual approach based on a language model pre-trained by Microsoft,[16] making it possible to incorporate the context of the word, and thus distinguish between "rate" where it is preceded by "interest" or "exchange". The algorithm used for this latter representation is capable of processing around 50 different languages.

[13] Analogies Explained: Towards Understanding Word Embeddings, Carl Allen and Timothy Hospedales

[14] The examples are regularly given in English, but the principles are the same for most languages, including French.

[15] A document is converted into a list of numbers of the size of the vocabulary in the corpus; each item in the list contains the number of occurrences of the word, or 0 if the word is absent.

[16] This is the multilingual MPNet model (link) retrained to identify insofar as possible expressions that are semantically similar.

2.3. RESULTS

The experiments which made it possible to develop the tool described in this report were conducted on a set of 171 annual financial documents taken from the sectors[17] of financial services (9 issuers), banking (7 issuers) and insurance (4 issuers) between 2012 and 2018, and on a number of URDs available in PDF format from 2019. However, given the problems faced concerning the ease of processing the latest URDs published in 2022 in XHTML format, these URDs were not used[18] for this study.

The display of the results offered by the tool can facilitate analyses in six different areas described in Table 2 below. A set of filters is proposed on each page to allow the teams to restrict the scope of the chart according to the selected parameters (issuer, sector, sub-sector, year, etc.).

Table 2: List of areas of analysis made possible by the display of results

Title	Description
Risk distributions	Investigation of risk proportions by issuer, year, super-sector, sector or sub-sector.
Change over time	Investigation of changes over time for each risk depending on the selected issuer, super-sector, sector or sub-sector. This page shows the appearance or disappearance of a risk, and its preponderance according to the selected sector.
Risk descriptions	Analysis of each risk identified during the post-processing phase; for each risk, it is possible to trace the main paragraphs according to the selected issuer and year.
Sector divergences	Alert system for presenting documents diverging furthest from the average risk proportions for a given sector. The documents diverging furthest are reported with an indication concerning the risk accounting for the over- or under-representation.
Divergences over time	Alert system making it possible to trace a document when the description of a risk for a given issuer has changed significantly in proportion relative to the previous year.
Comparison by issuer	Comparison of risk distributions from one issuer to another for a selected year, with the capability for reading paragraphs of interest when a risk is selected.

[17] Based on the international classification ICB “Industry Classification Benchmark” proposed by FTSE Group and Dow Jones Indexes.

[18] Yet, documents in this format are practically unreadable by machine. See: "Focus on the implications of an optimal use of machine-readable formats".

Figure 3 below shows, for example, the risk distribution obtained in the sample sub-sectors, and the average significance of the risks identified for a bank issuer in accounting year 2020. The pandemic risk, in salmon colour, is highlighted.

Figure 3: Risk distribution

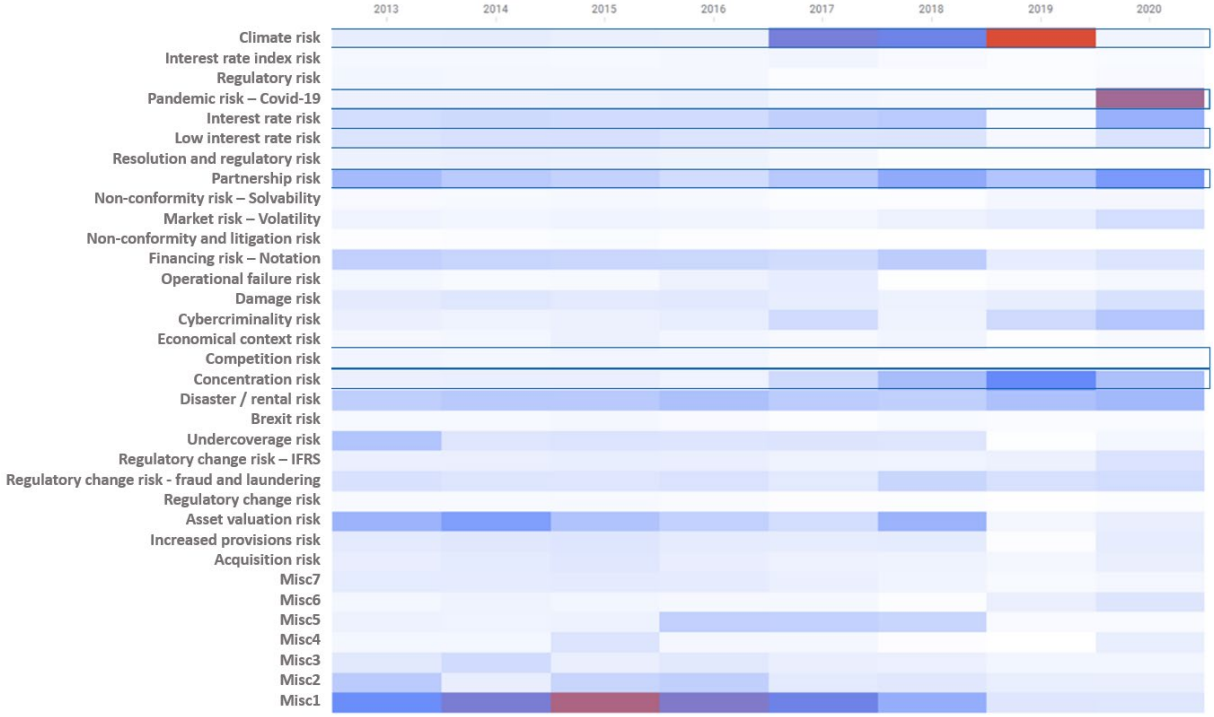
NB: Misc (Miscellaneous) corresponds to the markers identifying the themes for which no risk stands out in particular.

The above screen shot of the tool shows:

- in the top chart: average risk distributions for accounting year 2020 concerning the issuers of the sample grouped by sub-sector. This highlights, for example, the preponderance of the pandemic risk (cf. the size of the highlighted salmon-colour bars) for the life insurance and reinsurance sub-sectors, while it is mentioned far less by issuers in miscellaneous financial services and investment services.
- in the bottom chart: proportions by risk for a selected issuer, in this case a bank issuer for which the pandemic risk is in third position and concentration risk (cf. marine-blue-colour bars) fourth.

The results of the model also make it possible to explore "automatically" the change in risks over time. The following screen shot gives an idea of the change in mentions of risks each year on a selected sample (in this case insurers): the more the colour tends toward red, the more the risk is mentioned.

Figure 4: Change in risk factors, particularly pandemic risks, for the insurance sector



NB[1]: Misc (Miscellaneous) corresponds to the markers identifying the themes for which no risk stands out in particular.
NB[2]: For the year 2020, certain documents were not able to be processed[19] at the time of the experiments and this has an impact on the distributions, making it harder to interpret the results for that year.

In the above example, the mentions of pandemic risk rose dramatically in 2020. It is worth noting that this risk was not absent from the documents prior to the Covid-19 pandemic. This risk has been described since 2013 following the SARS epidemic, before decreasing from 2016 to 2019.

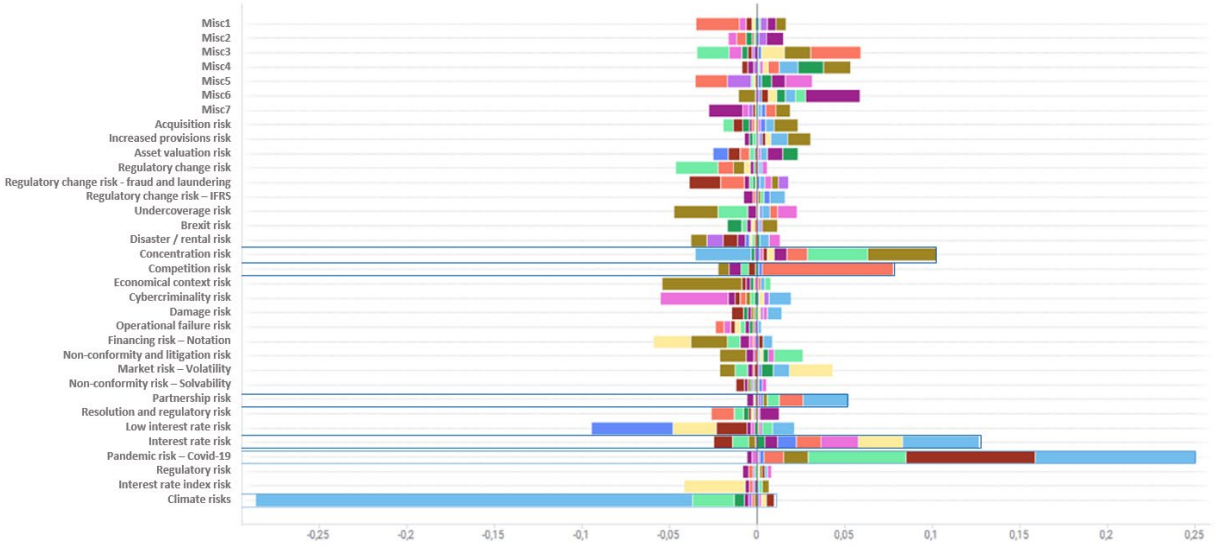
In terms of trends, an increase in interest-rate risk (excluding an anomaly in 2019) can be observed, as well as a sharp rise in concentration risk and an extremely significant partnership risk.

Moreover, whereas climate risk increased significantly from 2017 to 2019, with a sharp spike in the last year, it is surprising to see that the mentions of it decrease in 2020 (the phenomenon is explained below by Figure 5).

[19] These are documents provided by the issuer in XHTML format (see: "Focus on the implications of a machine-readable format"); only PDF documents are covered by the analysis.

Finally, because from one year to the next it is possible that an issuer may significantly change the risks that it describes, e.g. by reducing the magnitude of a risk that seems to it less substantial in the new year, or vice versa, the tool developed can also highlight these variations from one year to the next (see Figure 5 below).

Figure 5: Variation in time of the risk factors reported by the marketplace for accounting year 2020 compared with the previous year



NB: Misc (Miscellaneous) corresponds to the markers identifying the themes for which no risk stands out in particular.

Here, where each horizontal bar represents the change in a risk for all the issuers (each issuer has its own colour and can appear on several lines), it can be seen that between 2019 and 2020 (accounting years), an issuer in the insurance sector (light blue) significantly reduced the description of its exposure to climate risk. In this case, the change can be explained by this issuer's formatting of the document. Whereas, in accounting year 2019, this issuer devoted a large section to climate risk in the "Risk Factors" section, for 2020 it had shifted the portion relating to climate risk outside that section.

The increase in competition risk for a financial services issuer, here in orange, is also noteworthy, and is due to the addition of the risk of competitive pressure on management fee rates (whereas this risk was barely discussed in this issuer's previous reports).

More generally, moreover, it can be noted that:

- interest-rate risk increases for nearly all the issuers of the sample, and
- concentration and partnership risks are a matter of greater concern (as already seen via Figure 4 above for the insurance sector).

3. ADDITIONAL PATHS OF INVESTIGATION

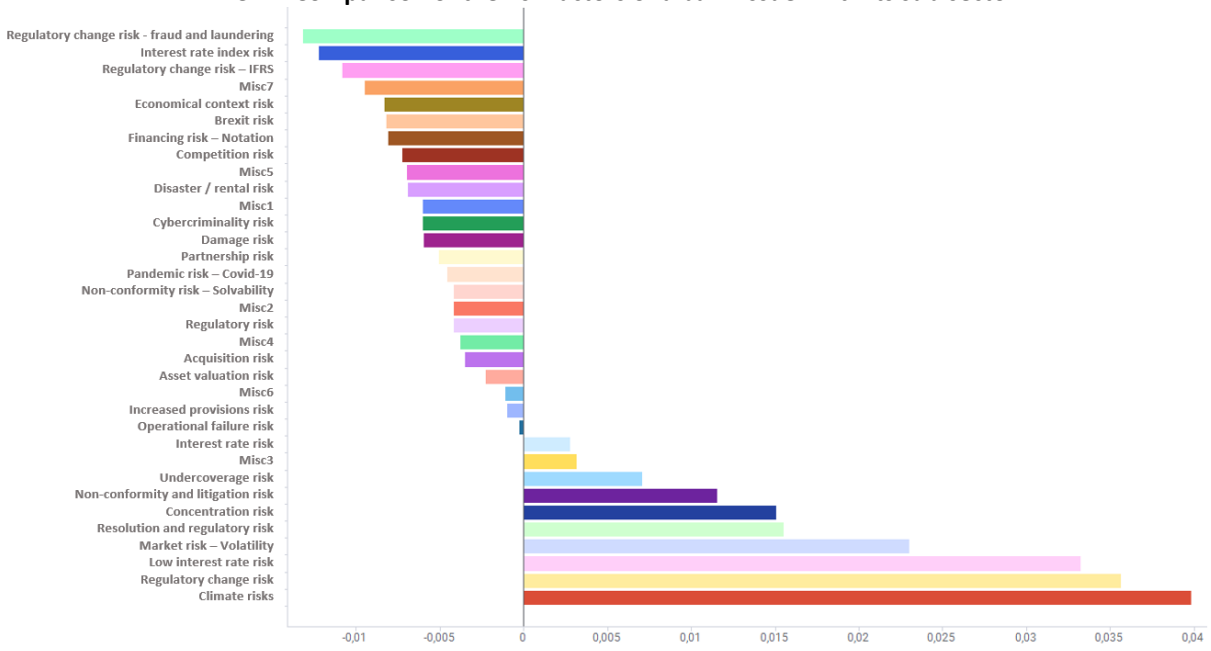
During these experimentations, a number of perspectives to be investigated in greater detail in the course of future experiments were identified, such as the generalisation to all the issuers of the French market, the analysis of regulatory constraints regarding the description of the specificity and materiality of risks, and the multilingual extension of the models to be able to process the documents of issuers throughout Europe from a comparative viewpoint.

***

Annex 1: Extract on the risk factor relating to the low-interest-rate environment for a bank issuer

"During periods of low interest rates, interest rate spreads tend to tighten, and the [HIDDEN] Group may be unable to lower interest rates on deposits sufficiently to offset reduced income from lending at lower interest rates. Net interest income amounted to EUR 21,127 million in 2019 and EUR 21,312 million in 2020, respectively. On an indicative basis, over one-, two- and three-year timeframes, the sensitivity of revenues at 31 December 2020 to a parallel, instantaneous and definitive increase in market rates of +50 basis points (+0.5%) across all currencies has an impact of +EUR 125 million, +EUR 309 million and +EUR 600 million, respectively, or +0.3%, +0.7% and +1.4% of the Group's net banking income. The negative interest rate environment, in which banks are charged for cash deposited with central banks, whereas banks typically do not charge clients for deposits, weighs significantly on banks' margins. In addition, the [HIDDEN] Group has been facing and may continue to face an increase in early repayment and refinancing of mortgages and other fixed-rate consumer and corporate loans as clients take advantage of lower borrowing costs."

Annex 2: Comparison of the risk factors of a bank issuer with its sub-sector



This chart presents the divergences in the presence of a risk for an issuer compared with its sub-sector. It provides a view of the relative significance of each of the firm's risks by comparison with its peers, thus providing greater depth for a comparative sector analysis and making it possible to quickly single out anomalies of over- or under-representation.

For example, in this case we have a bank issuer presenting in 2019 a climate risk on average 4 percentage points higher than the other issuers in its group. Also, the issuer seems to present little "Risk of regulatory change – Fraud and money laundering" but to be especially prolix regarding "Climate risk", "Risk of regulatory change", etc.

Annex 3: Theme detection models

Several types of "Theme Detection" models were experimented:

- matrix (Non-Negative Matrix Factorization, or NMF,[20] Latent Semantic Analysis,[21] or LSA);
- probabilistic (Latent Dirichlet Allocation,[22] or LDA), and
- neural for the state of the art ("Prod-LDA",[23] "SCHOLAR",[24] "Contextualized Topic Model",[25] or CTM, and "Covariate Zero-shot CTM").

The SCHOLAR model extends the capacity of Prod-LDA to be able to add to it (non-exhaustive list): metadata such as the sector to which the issuer belongs, and pre-trained word embedding. The benefit of adding metadata such as the sector or industry is to allow the model to pay more attention to the risks than to their variations from one sector to another (e.g. so that competition risk may be clearly identified as such, whether the issuer be from the banking or insurance industry).
Covariate Zero-shot CTM was built specifically for the project based on SCHOLAR and a variant of CTM[26] to extend the tool beyond the French language and thus explore the risk factors present in a maximum of European languages.

[20] Positive Matrix Factorization: A non-negative factor model with optimal utilization of error estimates of data values (Paatero et al., Environmetrics 1994)

[21] Latent Semantic Indexing: A Probabilistic Analysis (Papadimitriou et al., Journal of Computer and System Sciences, 2000)

[22] Latent Dirichlet Allocation (Blei et al., Journal of Machine Learning, 2003)

[23] Autoencoding Variational Inference For Topic Models (Srivastava et al., ICRL 2017)

[24] Neural Models for Documents with Metadata (Card et al., ACL 2018)

[25] Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence (Bianchi et al., ACL 2021)

[26] Cross-lingual Contextualized Topic Models with Zero-shot Learning (Bianchi et al., EACL 2021)