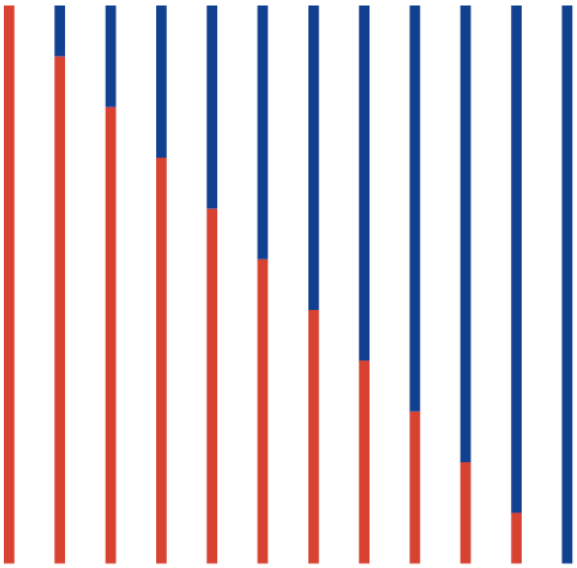


JANVIER 2023

ANALYSE AUTOMATIQUE DES FACTEURS DE RISQUES PUBLIES PAR LES SOCIETES COTEES : UN CAS D'USAGE DU TRAITEMENT DU LANGAGE NATUREL POUR L'AMF

CORENTIN MASSON

RÉSUMÉ

L'entrée en vigueur du règlement Prospectus en 2019 renforce les obligations de communication des émetteurs sur leur exposition aux risques à l'occasion de la publication de leur document d'enregistrement universel (DEU, ou URD en anglais). Dans la mesure où il incombe à l'AMF de veiller au bon respect des exigences en matière de présentation des risques, le régulateur a souhaité explorer les potentialités offertes par les technologies de Traitement Automatique du Langage Naturel (TAL, ou NLP en anglais, une branche de l'Intelligence Artificielle) afin d'analyser plus efficacement ces documents.

Les travaux de l'AMF se sont portés sur l'application de récentes avancées en apprentissage profond (i.e. d*eep learning* en anglais) susceptibles notamment de gérer les complexités sémantiques des sections des facteurs de risque telles que l'imbrication entre certains événements (par exemple, un risque géopolitique peut entraîner un risque de la flambée des prix énergétiques) et leurs déclinaisons spécifiques dans chacun des secteurs d'activité. A terme, l'usage de ces techniques permettrait également à l'AMF de faciliter la comparaison des risques des émetteurs de diverses nationalités en gérant un corpus de documents multilingue.

Sur la base de plus d'une centaine d'URD d'établissements financiers entre les années comptables 2012 à 2020, les expérimentations conduites montrent notamment qu'il est possible, de façon automatique, d'appréhender la répartition des risques par secteur ou par émetteur, de suivre leur évolution dans le temps. Il est aussi possible de détecter les variations les plus importantes d'une année sur l'autre dans les degrés de mention de chacun des risques présentés. A titre d'illustration, l'outil développé met visuellement en évidence l'émergence de la présentation du risque pandémique dans les URD publiés en 2021 (sur l'année comptable 2020) ainsi que la croissance continue du risque de sécurité informatique ou des hausses temporaires des risques réglementaires dans certains secteurs (par exemple liés au règlement Benchmark ou encore la loi de séparation des activités bancaires en 2013).

A l'issue de cette première phase d'expérimentation, mais sans toutefois minimiser les difficultés que représenterait la généralisation de cet outil à tout émetteur du marché français, les premiers résultats obtenus semblent prometteurs et permettent d'envisager à terme un recours aux techniques TAL plus largement pour soutenir les travaux du régulateur : à la fois dans le cadre de son suivi de l'information communiquée au marché par les émetteurs (et dans l'appréhension de la qualité de celle-ci), mais également pour alimenter la production d'études thématiques.

Enfin, pour faciliter l'exploitation automatique des documents émetteurs ou plus largement celle des documents réglementaires, l'AMF souhaite promouvoir l'usage des formats « *machine-readable* » et, surtout, l'associer à des bonnes pratiques permettant d'en optimiser l'utilisation.

1. LES ENJEUX DE L'EXTRACTION AUTOMATIQUE DES FACTEURS DE RISQUE POUR L'AMF

L'AMF, en son rôle de régulateur des sociétés cotées (alternativement désignées ci-après par « émetteurs » de titres cotés), veille à la qualité de l'information financière et extra-financière diffusée par ces sociétés dans le cadre de leurs obligations - périodiques et permanentes - et à l'occasion de leurs opérations financières.

L'information financière et extra-financière constitue un paramètre important dans la prise de décision des investisseurs sur les marchés. Parmi les nombreux éléments soumis à une obligation de communication de la part des émetteurs figure leur exposition aux risques évalués comme importants. Cette information se doit d'être précise et transparente afin qu'elle puisse contribuer à la décision d'investissement.

Depuis 2019, la réglementation européenne impose notamment un nouveau formalisme dans l'énonciation des facteurs de risque, en contraignant à la fois la forme et l'ordre de présentation et en ajoutant entre autres des critères d'importance et de spécificité[1] (Règlement (UE) 2017/1129 dit « Prospectus » venu abroger la directive (UE) 2003/71/CE). Dans le cadre de l'application de ce nouveau règlement, la charge de vérifier la conformité des sections « Facteurs de risque » revient à l'AMF.

Cette étude, aujourd'hui complètement manuelle, peut s'avérer longue et fastidieuse (plusieurs centaines de documents devant être analysés pour couvrir l'ensemble des émetteurs du marché français). Ainsi, le développement d'un outil automatisé d'analyse des facteurs de risque, englobant à la fois les techniques d'intelligence artificielle pour extraire les informations du texte et un ensemble de visualisations pertinentes des résultats de l'extraction, permettrait d'appuyer utilement les travaux conduits par l'AMF dans le cadre de ses missions.

Trois principaux intérêts ont été identifiés :

- les premières expérimentations ont montré qu'il était possible d'extraire les paragraphes les plus pertinents de chaque facteur de risque décrit par un émetteur, puis de les restituer aux équipes de l'AMF dans un unique écran pour leur consultation (exemple avec le risque de taux bas d'un émetteur en annexe 1). En proposant ensuite de consolider ces paragraphes dans une même visualisation pour l'ensemble des émetteurs d'un même secteur par exemple, cela permettrait aux équipes de **gagner du temps sur une première phase d'analyse qualitative** de la description des risques dans une optique de supervision.
- les tests conduits sur les modèles d'intelligence artificielle appliqués au traitement du texte des sections « Facteurs de risque » ont également mis en évidence la possibilité de **systématiser des analyses thématiques et de les étendre à l'ensemble des émetteurs** (voir ci-après la sous-section 1.3 Résultats), en particulier pour :
  - comparer par rapport à son secteur la manière dont un émetteur communique sur ses facteurs de risque (étude comparative sectorielle, *cf. figure 3 et annexe 2*) ;
  - conduire des analyses thématiques sur un type de risque donné et analyser ses modalités de présentation pour un ensemble d'émetteurs (par exemple les facteurs de risque climatiques au sein du secteur financier, étude détaillée par risque. *cf. figures 3 et 4 pour des exemples de graphiques pertinents pour une telle étude*).
  - appréhender l'évolution dans le temps des facteurs de risque d'un ou plusieurs émetteurs grâce à une visualisation aisée de tendances (*cf. figure 5 ci-après*), permettant d'identifier les facteurs de risque émergents d'un exercice à l'autre (tels que par exemple le risque associé à la crise sanitaire, *cf. figure 4 ci-après*) ;
- de futures explorations sur l'analyse automatisée des facteurs de risque pourraient offrir la possibilité de conduire des études de cohérence quant aux principaux éléments mis en exergue dans les différents supports de communication des émetteurs et de renforcer l'activité de veille sur la qualité de l'information transmise au marché : par exemple, en croisant les informations des émetteurs sur leurs facteurs de risque entre leurs documents réglementaires et celles qui sont transmises au moyen des

[1] Ces critères sont définis par l'Article 16 du Règlement Prospectus et sont précisés dans les guidelines de l'ESMA concernant les sections Facteurs de Risques, disponibles à cette adresse.

autres supports de communication financière et extra-financière (notamment les communiqués de presse).

Focus sur la recherche académique en finance utilisant du TAL

Bien que la diffusion d'informations dans la sphère financière est théorisée et étudiée depuis de nombreuses décennies, l'exploitation (incluant la lecture et l'analyse) automatique de documents a été plus tardive. Depuis les premières expérimentations au début des années 2000, nous identifions trois axes principaux de recherches, qui se distinguent principalement par les questions auxquelles ils cherchent à répondre.

Le premier inclut les travaux dont l'objectif principal est d'exploiter des données extraites automatiquement de documents textuels (journaux, annonces de résultats, documents régulés, etc.) afin d'étudier la manière dont les informations communiquées se diffusent dans la sphère financière, altèrent la perception des investisseurs ou peuvent nous renseigner sur les informations incluses dans les prix. Il s'agit d'obtenir une meilleure compréhension de la relation entre l'économie réelle et la sphère financière en étudiant le contenu des informations disponibles en langage naturel. Par exemple, en quantifiant automatiquement les primes de risques d'exposition au changement climatique, des chercheurs montrent que ces primes incombent plus à des chocs d'opportunités qu'à des risques physiques ou de régulation ; d'autres observent que détailler spécifiquement les risques dans les documents régulés réduit l'ampleur des chocs futurs. Le deuxième axe est très lié au premier mais se concentre sur l'intégration d'informations extraites des données textuelles dans les modélisations d'optimisation de portefeuille, de prédiction de volatilité ou de d'évaluation d'actifs. Par exemple, Engle et al. (2019) présentent une approche permettant de se protéger contre le risque climatique à partir de *news*. J. Lu et X. Huang (2021) proposent quant à eux d'utiliser de la détection d'événements pour prédire le prix du pétrole brut. Le dernier axe s'attache à analyser la qualité du contenu textuel en s'intéressant particulièrement aux documents régulés. On y retrouve, entre autres, des travaux sur la lisibilité des textes, leurs tons et autres techniques pour rendre le contenu plus favorable à l'émetteur qui le présente.

Il est à noter que l'analyse textuelle des facteurs de risque est particulièrement riche sur les rapports annuels soumis à la SEC (*10-K filings*) pour lesquels il existe une très importante masse de données aisément accessible[2] dans un format très exploitable par les machines[3]. En revanche, depuis le début de l'année 2022, les URD français dorénavant publiés sous format XHTML sont bien moins exploitables en dehors des sections sujettes au « format électronique unique européen » (en anglais *European Single Electronic Format*, ou ESEF)[4] (cf. : « Focus sur les enjeux d'une utilisation optimale des formats *machine-readable*»).

Enfin, de nombreuses publications académiques[5] existent sur l'influence de la présentation des risques sur les rendements boursiers d'un émetteur, sur la manière dont sont perçus et exploités ces risques par les investisseurs, ou encore sur les gains de performance que peuvent offrir ces données pour protéger un portefeuille de certains risques identifiés.

[2] L'ensemble des documents est accessible via la plateforme EDGAR sur laquelle il est possible de chercher manuellement ou automatiquement via une API (Application Programming Interface).

[3] Inline XBRL, XHTML avec une contrainte de standardisation s'appliquant à l'ensemble du document.

[4] Cf. : présentation des enjeux d'ESEF sur le site de l'ESMA.

[5] Telles que : " 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 sur les enjeux d'une utilisation optimale des formats *machine-readable*

Dans la directive (EU) 2019/1024 dite « Open-Data »[6], la Commission Européenne définit quels sont les formats de document qui permettent leur exploitation automatique par une machine (en anglais « *machine-readable format* »), et promeut leur utilisation afin de faciliter l'accès aux données[7]. Parmi ces formats, le « XHTML » permet par exemple la rédaction d'un document avec l'usage d'un système de balises pour définir à la fois la structure du contenu (titre, section, sous-section, etc.) et le référencement à certaines informations spécifiques à l'intérieur du texte (ou d'un tableau).

Si les réglementations s'attachent surtout à définir les informations clés qui doivent être balisées (notamment dans le cadre ESEF), en pratique, dans le cas du recours au format XHTML, l'emploi des balises relevant davantage de la structure du document est capital pour permettre l'exploitation du document dans sa globalité. Par exemple, les balises définissant les sections ou les niveaux de titres permettent à une machine de naviguer plus facilement dans un document de plusieurs centaines de pages afin d'isoler une portion du document, comme une section en particulier. De la même façon, le format XHTML prévoit des balises pour la structure des tableaux dont l'usage permet de faciliter l'extraction des données sans avoir besoin de recourir à des techniques sophistiquées comme de la reconnaissance d'image.

Aussi, la seule utilisation du format XHTML n'est pas suffisante pour que les documents soient en pratique exploitables par la machine dans leur intégralité, c'est-à-dire au-delà des informations clés balisées. Ainsi, les tentatives d'utilisation des URD 2022 publiés au format XHTML pour l'analyse des facteurs de risques (qui ne sont pas des informations clés balisées en application de la réglementation) n'ont pas été concluantes du fait notamment de l'absence de balise pour la structure du document ou la construction des tableaux et de l'utilisation des balises pour le placement des mots dans le document[8]. A cause de ces traitements non pertinents, repérer une section particulière du document et traiter les tableaux est très complexe. Encore plus dommageable, les mots et paragraphes peuvent apparaître dans le désordre par rapport au rendu visuel et des mots peuvent être coupés ou fusionnés avec d'autres ce qui rend le contenu de certains documents presque inutilisable[9].

L'AMF souhaite attirer l'attention des producteurs de documents sur l'importance de suivre des bonnes pratiques pour optimiser la qualité des fichiers XHTML, en particulier : l'utilisation des balises appropriées pour séparer les sections et les paragraphes ainsi que pour structurer les tableaux, et le bannissement des pratiques qui permettent d'intervertir l'ordre des phrases ou des mots dans le code par rapport au rendu visuel.

2. EXPÉRIMENTATIONS : APPROCHES ET RÉSULTATS

2.1. PRÉSENTATION DES RISQUES PAR LES ÉMETTEURS

Dans la section des URD dédiée aux « Facteurs de risque », les émetteurs décrivent les différents risques auxquels ils sont exposés comme un ensemble d'événements aléatoires pouvant avoir un impact négatif important sur leurs résultats ou leur croissance.

Il existe de nombreux types de risques, tels que les risques climatiques ou les risques réglementaires, qui peuvent eux-mêmes être déclinés en plusieurs risques distincts : par exemple dans son rapport 2020, un assureur présentait le risque d'évolution réglementaire sous cinq aspects, et notamment : « Exigences en termes de fonds-propres », « Enjeux liés au blanchiment d'argent et à la corruption », « Réforme des benchmarks » ou encore « Changement des normes IFRS ». En pratique, chacune de ces variantes constitue un risque à part entière.

---

[6] Article 2, paragraphe 13

[7] La promotion de ce type de format se retrouve aussi dans les travaux de l'EFRAG relatifs aux normes européennes de rapport sur le développement durable (ESRS), et plus particulièrement dans ses propositions d'exigences générales, voir *DRAFT ESRS 1 General Requirements* (lien).

[8] Ce qui engendre de nombreuses erreurs dans l'extraction de texte, telles que l'inversion d'ordres des paragraphes, des mots et/ou des lettres, la concaténation ou la séparation de certains mots et des disparitions de lettres.

[9] Ces défauts de qualité proviennent de la conversion de documents Word au format XHTML par des logiciels tierces, qui n'utilisent pas les balises pour leurs rôles prévus par les standards HTML et XHTML du *World Wide Web Consortium* (lien).

Par ailleurs, les émetteurs ne se contentent pas d'énumérer distinctement les facteurs de risque les uns après les autres, mais mettent également en avant une très forte imbrication des risques entre eux. Aussi, une hausse des taux peut entraîner un risque de crédit, ou un risque géopolitique peut entraîner un risque de la flambée des prix énergétiques (cf. ci-dessous « Focus sur un paragraphe de risque dans l'URD 2020 d'un assureur » pour un exemple de paragraphe de risque incorporant des éléments de plusieurs risques imbriqués).

Enfin, si les orientations de l'ESMA précisent la manière dont les exigences réglementaires relatives aux facteurs de risque doivent être mises en œuvre, en pratique, il peut être constaté que les risques sont explicités de différentes manières, notamment :

- soit pour tenir compte des particularités du secteur (par exemple, les risques opérationnels ou concurrentiels peuvent s'avérer de natures très différentes selon si l'émetteur est une banque, un assureur ou une société de travaux publics),
- soit dans leurs niveaux de détails (dans son URD portant sur l'exercice 2020 une banque décrivait un risque de crédit relatif à ses activités de prêt et un risque de crédit relatif à sa détention d'obligations, tandis qu'un assureur distinguait deux risques de crédit, un premier sur les obligations privées et un second sur les obligations souveraines).

Aussi, si un analyste saura interpréter sans trop de difficultés les nombreuses informations de cette section, automatiser l'analyse par une machine constitue un véritable défi.

Focus sur un paragraphe de risque dans l'URD 2020 d'un assureur.

« Les résultats du Groupe pourraient être affectés de manière significative par la situation économique et financière en Europe et dans d'autres pays du monde. La menace d'une dépression économique mondiale due à des facteurs sanitaires, cycliques et/ou commerciaux (par exemple, l'actuelle guerre commerciale entre la Chine et les États-Unis) demeure, et une détérioration macroéconomique durable pourrait affecter les activités et les résultats de la société. Les taux d'intérêt ont atteint un niveau historiquement bas, et s'ils devaient augmenter, les niveaux exceptionnels actuels de l'endettement public et privé pourraient devenir source d'une instabilité financière majeure. Également, tout assouplissement supplémentaire de la politique monétaire aurait peu d'effets sur l'économie. Ces tendances pourraient entraîner une période de très forte volatilité sur les marchés financiers, pouvant engendrer une vague de faillites d'entreprises et potentiellement de défauts souverains dans les régions vulnérables, une chute de la valeur des principales classes d'actifs (obligations, actions, immobilier), ou une crise majeure de liquidité. En l'absence d'un déploiement rapide et massif des vaccins contre le Covid-19 dans la population, les perspectives économiques restent difficiles. De plus, les difficultés économiques actuelles des États-Unis et les disparités économiques persistantes entre les pays européens pourraient contribuer à d'autres impacts politiques et économiques. Pour plus d'informations sur le portefeuille de placement de la société, se référer aux sections 1.3.9.2 – Le rendement sur investissements et le rendement sur actifs investis et 4.6 – Annexe aux comptes consolidés, note 8 – Placements des activités d'assurance. »

Ce paragraphe présent dans la sous-section « Risques liés à l'environnement macroéconomique » et plus particulièrement la portion « Risques de détérioration des marchés financiers et de l'économie mondiale » est un exemple type de l'imbrication des risques. L'émetteur y dépeint de manière qualitative les tensions pesant sur le contexte économique : les risques liés au contexte économique (bleu), de hausse des taux d'intérêt (jaune), de défaut (orange), de continuité de la pandémie de Covid-19 (vert) et de liquidité (violet). Chacun de ces facteurs peut dériver en risque de prix (rouge).

2.2. L'APPROCHE RETENUE POUR AUTOMATISER L'ANALYSE DES RISQUES PAR UNE MACHINE

Afin d'aider l'AMF dans ses missions d'analyse des facteurs de risque, l'outil doit tout d'abord être en mesure :

- d'identifier les principaux risques présents dans chacun des documents publiés par les émetteurs,
- d'estimer l'importance de chacun de ces risques, et,
- de référencer les paragraphes du document associés à chaque risque.

Les URD sont des documents très denses (plusieurs centaines de pages) contenant diverses sections dont celle sur les facteurs de risque (plus d'une dizaine de pages en moyenne). La machine doit donc être capable

d'identifier les portions précises du document discutant des risques. Pour cela, une première brique technique[10] a été développée afin d'identifier les paragraphes et pages pertinents[11].

Ensuite, parmi les paragraphes conservés à la suite de cette première étape, l'outil doit distinguer les différents types de risques qui y sont cités. Pour cela, chaque risque peut être vu comme « un thème », caractérisé par la présence d'un ensemble de mots appartenant au même champ lexical. La « détection de thèmes » est une tâche courante des travaux en TAL, et de nombreux modèles[12], fondés sur la recherche des champs lexicaux, ont déjà été proposés dans la littérature académique. Sur la base des modèles existants, un algorithme spécifique a été développé pour l'analyse des facteurs de risque dans lequel chaque thème est ensuite associé à un risque (voir tableau 1 ci-après).

Tableau 1: Exemples de risques identifiés

Champs lexical	**Risque associé**
tentative – informatique - intrusion - confidentiel - cyber - attaque - malveillant - piratage - obsolescence - cyberattaque	« Risque de cybercriminalité »
transition – investissement* - empreinte – changement – charbon - climatique - environnemental - hydrocarbure – carbone - esg	« Risque climatique »
propagation – incertitude – mondial – mesure – naturel – apparition – transmission – virus – coronavirus - vague	« Risque pandémique »
taux - variation – investissement* - devise - fluctuation - duration - change - valeur - obligataire - rendement	« Risque de taux d'intérêt / de change »
amende - loi - contentieux - divergent - avertissement – annuel* - applicable - constant - sanction - correctif - code - texte - objet - adoption - voire*	« Risque de non-conformité »

*NB : Les mots annotés par un astérisque « * » dans le tableau ci-dessus ne sont pas spécifiquement liés au champ lexical du risque auquel ils sont rattachés. En effet, le modèle utilisé se base sur des statistiques d'apparitions conjointes de mots afin d'identifier les thèmes, dont certaines ne sont pas toujours pertinentes (comme par exemple « voire » et « annuel » qui apparaissent fréquemment dans les paragraphes relatifs au « Risque de non-conformité », et « investissement » qui apparait à la fois dans ceux des « Risques climatiques » et dans les « Risques de taux » mais pas dans les autres).*

En pratique, l'algorithme renverra pour chaque paragraphe analysé les thèmes détectés (autrement dit les risques ; voir tableau 1 ci-dessus) avec une probabilité de confiance. Etant donné que l'étude se restreint pour le moment aux documents en langue française, et après une évaluation détaillée des performances, le modèle de « détection de thèmes » qui a finalement été implémenté est un modèle flexible et à l'Etat de l'Art particulièrement adapté à notre problématique. Ce modèle, nommé « SCHOLAR », a été paramétré, entraîné et validé sur nos données.

Enfin, la machine doit analyser les informations obtenues sur les paragraphes pour estimer quelle importance est donnée à chaque risque (i.e. thème). En pratique, cela signifie notamment de savoir associer deux (ou plus) informations entre elles, sans qu'elles soient nécessairement proches les unes des autres dans le texte. Afin d'obtenir une répartition du poids des risques mentionnés par un émetteur, l'outil va comparer la taille des paragraphes consacrés à la description de chacun. Autrement dit, il est considéré que l'importance d'un risque est directement proportionnelle à la quantité de texte qui lui est consacrée.

[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.

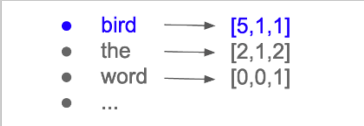
[11] En 2022, des travaux sont également en cours pour reconstruire le sommaire de chaque document, sur la base de la compétition FinTOC proposée dans le cadre du *Workshop Financial Narratives processing* co-localisé à LREC 2022 (lien).

[12] Voir annexe 3 pour plus de détails sur les modèles de "Détection de Thèmes".

Focus « Comment la machine traite le texte ? »

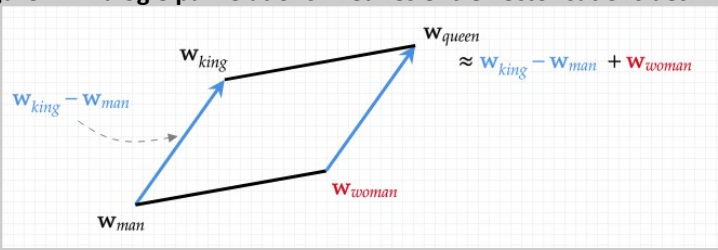
En tant que chaîne de caractères, de mots et de phrases, un texte n'est pas directement exploitable par une machine : une transformation numérique préalable est nécessaire, appelée communément « vectorisation » (cf. Figure 1 ci-dessous).

Figure 1: Vectorisation ou transformation numérique des mots



Il existe différentes approches pour procéder à cette transformation numérique et les approches sophistiquées permettent de conserver plusieurs propriétés de sens et de syntaxe. Par exemple, il a été démontré[13] que certains de ces vecteurs numériques identifient des analogies par des relations linéaires : les mots anglais « *king*, « *woman* » et « *man* » permettent de calculer le vecteur représentant le mot « *queen* »[14] (cf. figure 2 ci-dessous).

Figure 2: Analogie par relations linéaires entre vectorisations des mots



Trois approches de vectorisation ont été étudiées pour les travaux sur l'analyse des facteurs de risque :

- en « sacs-de-mots »[15] : dans sa version la plus simple, le texte transformé est alors représenté par l'histogramme des occurrences des mots le composant ;

- en « plongements lexicaux non contextuels » basée sur l'algorithme « skip-gram » : chaque mot est représenté par un vecteur et la transformation numérique doit permettre de conserver l'information que deux mots sont proches de sens ;

- en « plongements lexicaux contextuels » : une approche multilingue basée sur un modèle de langue pré-entraîné par Microsoft[16] et permettant d'incorporer le contexte du mot, et donc de différencier « taux » dans le cas où il est suivi d' « intérêt » ou de « change ». L'algorithme utilisé pour cette dernière représentation est capable de traiter près de 50 langues différentes.

---

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

[14] Les exemples sont régulièrement donnés en anglais, mais les principes sont les mêmes pour la majorité des langues, y compris pour le français.

[15] Un document est transformé en une liste de chiffres de la taille du vocabulaire du corpus, chaque élément de la liste contient le nombre d'occurrences du mot ou 0 si celui-ci est absent.

[16] Il s'agit du modèle MPNet multilingue (lien) ré-entraîné afin d'identifier au mieux les phrases proches sémantiquement.

2.3. RÉSULTATS

Les expérimentations qui ont permis de développer l'outil présenté dans cette note ont été conduites sur un ensemble de 171 documents financiers annuels issus des secteurs[17] des services financiers (9 émetteurs), bancaire (7 émetteurs) et assuranciel (4 émetteurs) entre 2012 et 2018, ainsi que sur un certain nombre d'URD disponibles en format PDF à partir de 2019. En revanche, au regard des problèmes rencontrés sur l'exploitabilité des derniers URD publiés en 2022 sous format XHTML, ces derniers n'ont pas été exploités[18] pour cette étude.

La visualisation des résultats proposée par l'outil permet de faciliter les analyses sur six différents axes décrits dans le tableau 2 ci-dessous. Un ensemble de filtres est proposé sur chaque page afin de permettre aux équipes de restreindre le périmètre du graphique selon les paramètres choisis (émetteur, secteur, sous-secteur, année, etc.).

Tableau 2 : Liste des axes d'analyse permis par la visualisation des résultats

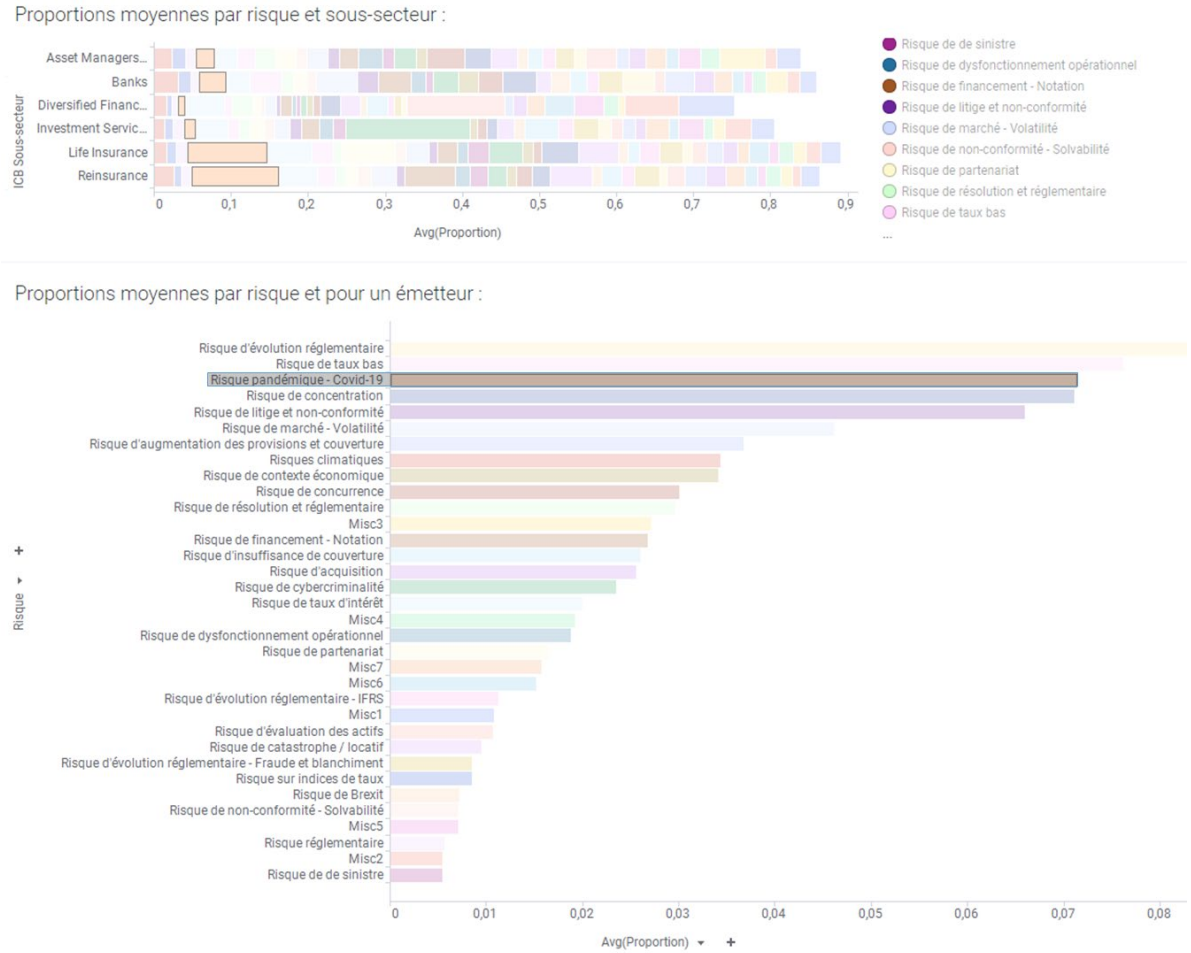
Titre	Description
Distributions de risques	Exploration des proportions des risques par émetteur, année, super-secteur, secteur ou sous-secteur.
Evolution temporelle	Exploration des évolutions temporelles de chaque risque selon l'émetteur, le super-secteur, le secteur ou le sous-secteur choisi. Cette page permet de voir l'apparition ou la disparition d'un risque, sa prépondérance selon le secteur choisi.
Descriptions des risques	Analyse de chaque risque identifié lors de la phase de post-traitement, il est possible pour chaque risque de remonter les paragraphes principaux selon l'émetteur et l'année choisi.
Divergences sectorielles	Système d'alerte pour présenter les documents s'éloignant le plus des proportions moyennes des risques pour un secteur donné. Les documents les plus éloignés sont remontés avec une indication sur le risque responsable de la sur ou sous-représentation.
Divergences temporelles	Système d'alerte permettant de remonter un document lorsque la description d'un risque pour un émetteur donné a significativement changé en termes de proportion par rapport à l'année précédente.
Comparaison par émetteur	Comparaison des distributions de risques d'un émetteur à un autre pour une année choisie, avec la capacité de lire les paragraphes d'intérêt lorsqu'un risque est choisi.

[17] D'après la nomenclature internationale ICB "Industry Classification Benchmark" proposée par FTSE Group et Dow Jones Indexes.

[18] A ce jour les documents dans ce format sont pratiquement illisibles par une machine, cf. : « Focus sur les enjeux d'une utilisation optimale des formats *machine-readable* ».

La figure 3 ci-dessous montre par exemple les distributions des risques obtenues dans les sous-secteurs de l'échantillon, ainsi que l'importance moyenne des risques identifiés pour un émetteur bancaire sur l'année comptable 2020. Le risque pandémique, couleur saumon, est mis en surbrillance.

Figure 3 : Distributions de risques



NB : Misc (Miscellaneous) correspond aux marqueurs identifiant les thèmes pour lesquels aucun risque ne se démarque particulièrement.

La capture d'écran de l'outil ci-dessus met en évidence :

- dans le graphique du haut : les distributions moyennes des risques pour l'année comptable 2020 sur les émetteurs de l'échantillon regroupés par sous-secteur. Celui-ci met par exemple en évidence la prépondérance du risque pandémique (cf. la taille des barres de couleur saumon en surbrillance) pour les sous-secteurs de l'assurance vie et de la réassurance alors qu'il est nettement moins mentionné chez les émetteurs des services financiers divers et des services d'investissements.
- dans le graphique du bas : les proportions par risque pour un émetteur sélectionné, ici, un émetteur bancaire pour lequel le risque pandémique est en troisième position et le risque de concentration (cf. barres de couleur bleu marine) en quatrième.

Les résultats du modèle permettent également d'explorer « automatiquement » l'évolution temporelle des risques. La capture d'écran ci-dessous permet d'apprécier l'évolution par année des mentions de risques sur un échantillon choisi (d'assureurs ici) : plus la couleur tire vers le rouge, plus le risque est mentionné.

Figure 4: Evolution des facteurs de risque, notamment pandémique sur le secteur assuranciel

NB[1] : Misc (Miscellaneous) correspond aux marqueurs identifiant les thèmes pour lesquels aucun risque ne se démarque particulièrement.
NB[2] : Sur l'année 2020 certains documents n'étaient pas exploitables[19] au moment de l'expérimentation, ce qui a un impact sur les distributions rendant plus difficilement interprétables les résultats sur cette année.

Dans l'exemple ci-dessus, les mentions du risque pandémique sont en hausse brutale en 2020. Il est intéressant de noter que ce risque n'était pas absent des documents avant la pandémie de Covid-19. En effet celui-ci est décrit depuis 2013 suite à l'épidemie de SARS avant de décroître de 2016 à 2019.

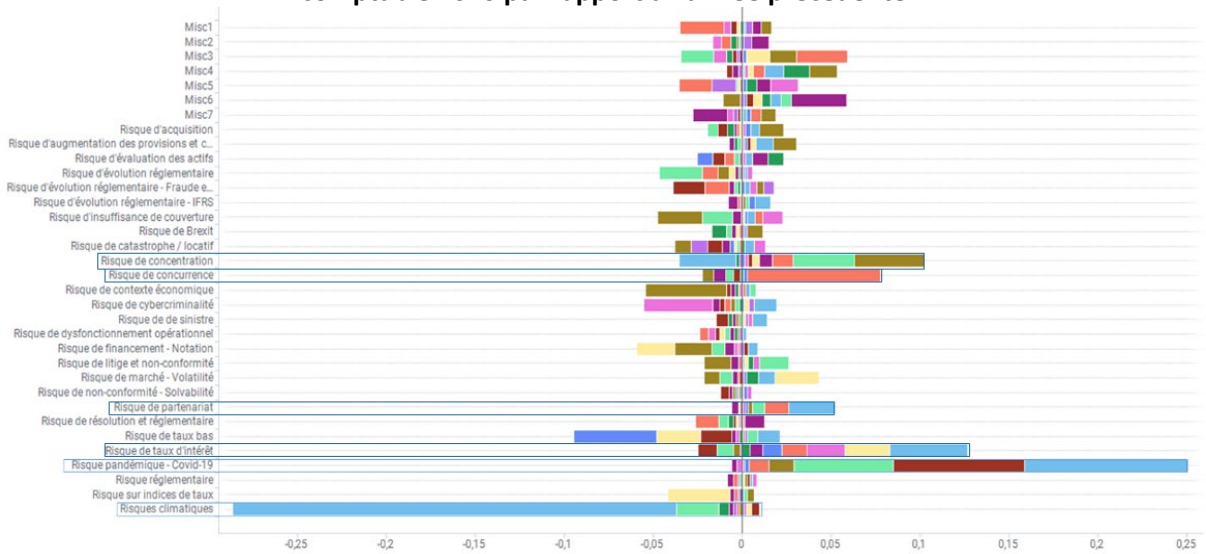
En tendance, un accroissement du risque de taux d'intérêt (hormis une anomalie en 2019) est observable, il en va de même pour une forte hausse du risque de concentration ainsi qu'un risque de partenariat particulièrement important.

Par ailleurs, alors que le risque climatique augmente fortement de 2017 à 2019 avec un fort pic sur la dernière année, il est étonnant de voir les mentions de celui-ci diminuer en 2020 (le phénomène est expliqué ci-après par la figure 5).

[19] Il s'agit ici des documents fournis par l'émetteur en format XHTML (cf. : « Focus sur les enjeux d'un format *machine-readable* »), seuls les documents PDF sont traités dans l'analyse.

Enfin, parce que d'une année sur l'autre il est possible qu'un émetteur change fortement les risques qu'il présente, par exemple en réduisant la taille d'un risque lui semblant être moins important sur cette nouvelle année ou inversement, l'outil développé permet également de mettre en évidence ces variations d'une année sur l'autre (voir figure 5 ci-dessous).

Figure 5: Variation temporelle des facteurs de risque communiqués par la place sur l'année comptable 2020 par rapport à l'année précédente



NB : Misc (Miscellaneous) correspond aux marqueurs identifiant les thèmes pour lesquels aucun risque ne se démarque particulièrement.

Ici, alors que chaque barre horizontale représente la variation d'un risque pour l'ensemble des émetteurs (chaque émetteur a sa propre couleur et peut figurer sur plusieurs lignes), il apparaît qu'entre 2019 et 2020 (années comptables), un émetteur du secteur des assurances (en bleu clair) diminue fortement la présentation de son exposition au risque climatique. En l'occurrence, cela s'explique par une question de mise en forme du document de cet émetteur. En effet, alors que, sur l'année comptable 2019, cet émetteur consacrait une large section relative aux risques climatiques au sein de la section des « Facteurs de risque », pour l'année 2020 cet assureur avait déplacé la portion relative aux risques climatiques en dehors de cette section.

L'augmentation du risque de concurrence pour un émetteur des services financiers, ici en orange, est également notable et provient de l'ajout du risque de pression concurrentielle sur les taux des commissions de gestion (alors que celui-ci était à peine discuté dans les rapports précédents de cet émetteur).

Par ailleurs, de façon plus générale, il peut être observé que :

- le risque de taux d'intérêt s'accroît pour la quasi-totalité des émetteurs de l'échantillon, et,
- les risques de concentration et de partenariat font l'objet d'une préoccupation plus importante (comme déjà observé via la figure 4 pour le secteur assurantiel ci-dessus).

3. PISTES D'EXPLORATION COMPLÉMENTAIRES

Durant ces expérimentations un certain nombre de perspectives à approfondir dans le cadre de futurs travaux ont été identifiés, telles que la généralisation à l'ensemble des émetteurs du marché français, l'analyse des contraintes réglementaires s'agissant de la présentation de la spécificité et de la matérialité des risques et l'extension multilingue des modèles pour permettre de traiter les documents des émetteurs au niveau européen dans une optique de comparaison.

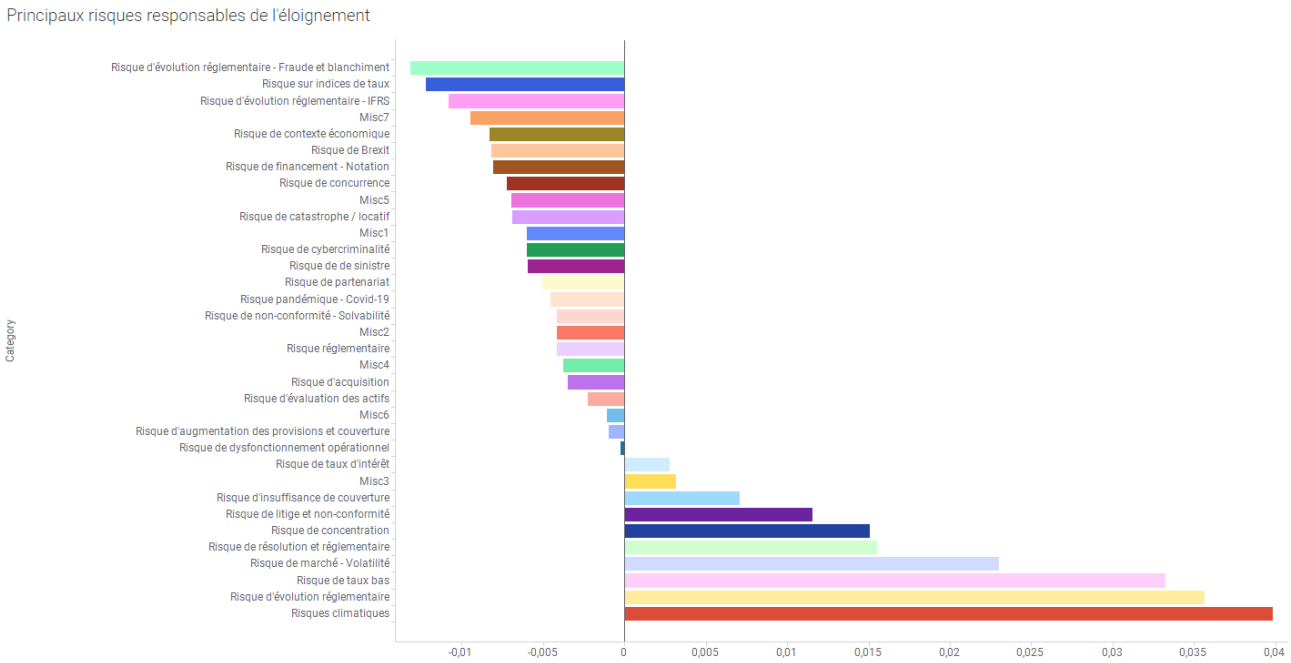
***

Annexe 1 : Extrait du facteur de risque relatif à l'environnement de taux d'intérêt bas pour un émetteur bancaire

Paragraphe choisi :

Durant les périodes de taux d'intérêt bas, les écarts de taux d'intérêt tendent à se resserrer ; le Groupe peut alors ne pas être en mesure d'abaisser suffisamment les taux d'intérêt sur ses dépôts de manière à compenser la baisse de revenus provenant des prêts consentis à des taux plus faibles. La marge d'intérêts s'élevait respectivement à 21 062 millions d'euros en 2018 et à 21 127 millions d'euros en 2019 (voir la note 3a « Marge d'intérêts » des états financiers consolidés). À titre indicatif, sur les horizons de un, deux et trois ans, la sensibilité des revenus au 31 décembre 2019 à une augmentation parallèle, instantanée et définitive des taux de marché sur l'ensemble des devises de + 50 points de base (+ 0,5 %) a un impact de respectivement - 270 millions d'euros, + 216 millions d'euros et + 614 millions d'euros ou - 0,6 %, + 0,5 % et + 1,4 % du produit net bancaire du Groupe. Un environnement de taux négatifs impliquant une facturation des liquidités déposées par les banques auprès des banques centrales alors que les dépôts bancaires ne sont usuellement pas facturés par les banques à leurs clients, constitue un facteur tendant à réduire la marge des établissements bancaires. De plus, le Groupe a fait et pourrait encore faire face à une hausse des remboursements anticipés et des refinancements de prêts hypothécaires et autres prêts à taux fixe consentis aux particuliers et aux entreprises, les clients cherchant à tirer parti de la baisse des coûts d'emprunt.

Annexe 2 : Comparaison des facteurs de risque d'un émetteur bancaire avec son sous-secteur



Ce graphique présente les écarts de présence d'un risque pour un émetteur comparé à son sous-secteur. Il permet de visualiser les importances relatives de chaque risque de l'acteur par rapport à ses pairs, offrant ainsi plus de profondeur à une analyse comparative sectorielle et la capacité d'isoler rapidement des anomalies de sur ou de sous-représentation.

Par exemple, il s'agit ici d'un émetteur bancaire présentant en 2019 un risque climatique en moyenne 4 points de pourcentage plus long que celui des autres émetteurs de son groupe. Aussi, l'émetteur semble peu présenter de « Risque d'évolution réglementaire – Fraude et blanchiment » mais être particulièrement profus sur les « Risque climatiques », « Risque d'évolution réglementaire », etc.

Annexe 3 : Les modèles de détection de thèmes

Plusieurs types de modèles de « Détection de Thèmes » ont été expérimentés :

- matriciels (*Non-Negative Matrix Factorization* ou NMF[20], *Latent Semantic Analysis*[21] ou LSA),
- probabiliste (*Latent Dirichlet Allocation*[22] ou LDA) et
- neuronaux pour l'état de l'art (*Prod-LDA*[23], *SCHOLAR*[24], *Contextualized Topic Model*[25] ou CTM et *Covariate Zero-shot CTM*).

Le modèle *SCHOLAR* étend les capacités du modèle *Prod-LDA* pour permettre d'y ajouter (non exhaustif) : des métadonnées telles que le secteur auquel appartient l'émetteur et des plongements lexicaux pré-entraînés. L'intérêt d'ajouter des métadonnées telles que le secteur ou l'industrie est de permettre au modèle de porter plus d'attention aux risques qu'à leurs variations d'un secteur à l'autre (par exemple pour que le risque de concurrence soit bien identifié comme tel, que l'émetteur soit de l'industrie bancaire ou assurancielle).
Covariate Zero-shot CTM a été construit spécifiquement pour le projet à partir de *SCHOLAR* et une variante de *CTM*[26] pour généraliser l'outil au-delà de la langue française et ainsi explorer les facteurs de risque présents dans un maximum de langues européennes.

[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).