Evaluating Topic Modeling on Imbalanced Multi-domain Financial Corpus for Risk Extraction

Corentin Masson[*†], **Patrick Paroubek**[†]
[*]AMF, [†]LISN-CNRS-U. Paris-Saclay
AMF 17 Place de la Bourse, 75002, Paris F
LISN Bât. 507, Rue John Von Neumann, 91400 Orsay F
c.masson@amf-france.org, patrick.paroubek@lisn.upsaclay.fr

Abstract

Multiple recent research works in Finance try to quantify the exposure of market assets to various risks from text and how assets react if the risk materialize itself. We consider risk sections from french Financial Corporate Annual Reports, which are regulated documents with a mandatory section containing important risks the company is facing, to extract an accurate risk profile and exposure of companies. We identify multiple pitfalls of topic models when applied to corporate filing financial domain data for unsupervised risk distribution extraction which has not yet been studied on this domain. We propose two new metrics to evaluate the behavior of different types of topic models with respect to pitfalls previously mentioned about document risk distribution extraction. Our evaluation will focus on three aspects: types of topic models, regularizations and down-sampling. In our experiments, we found that classic topic models require down-sampling to obtain unbiased risks, while topic models using metadata and in-domain pre-trained word-embeddings partially correct the coherence imbalance per subdomain and remove sector's specific language from the detected themes. We then demonstrate the relevance and usefulness of the extracted information with visualizations that help to understand the content of such corpus and its evolution along the years. Our conclusions are not restricted to the french language.

Keywords: Finance, Topic Modeling, Evaluation

1. Introduction

Today's economic context and perspective are highly volatile and uncertain, which is reflected on market prices and volatility (Taleb, 2007; Baker et al., 2016), with the result of a tendency for financing costs to increase. As exposures to probable material events are only partially priced on markets (Grossman and Stiglitz, 1980; Chung et al., 2012; Bao and Datta)) often caused by withheld or obfuscated information (Badawy and Ibrahim, 2016; de Souza et al., 2019), regulators ask for more precise and exhaustive non-financial disclosure about the risks companies are facing. New regulations appeared recently, such as the Prospectus Directive from the European Commission, *Sustainable Finance Disclosures Regulation* (SFDR) [1] for financial intermediaries and its counterpart *Corporate Sustainaility Reporting Directive* (CSRD) for listed companies, or the guidelines from the *Task Force on Climate-Related Financial Disclosures* (TCFD) already increases the amount and quality of publicly available textual data on companies' risks exposure.

As the financial academic literature demonstrate, investors as well as regulators show a growing interest in building systems capable of extracting information on companies risks and exposures from natural language documents, either regulated ones, such as "Risk Factors" sections in 10-K filings or Universal Registering Documents and Earning Calls or unregulated ones, such as News and social medias. We propose to expand and deepen existing work on unsupervised risk extraction from "Annual Report" types of documents on the French language and markets, which has not yet been studied. The majority of such works coming from the financial academic community, they rely on relatively simple Natural Language Processing systems, whose pitfalls have not been studied yet in detail.

We focus on such documents released each year and containing a description of companies specific and material risk factors, which we believe to be the most comprehensive indicator at a given time. Even if a great part of researchers work on Earning Calls for their immediate availability they tend to be biased towards the most current subjects. As an example, right before the Covid-19 pandemic this type of risk wasn't discussed in Earning Calls but was present in a significant part of Annual Reports (Loughran and McDonald). Documents in our corpus are limited to the Financials industry [2] in the french market, with Reference Documents (RD) from 2011 to 2018 and Universal Registering Documents (URD)[3] from 2019 to 2020. These documents have multiple characteristics that complexify unsupervised automatic pro-

[1] "https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014L0095"

[2] ICB Industry Classification

[3] Only if the document is available in a PDF format.

LREC-COLING 2024, pages 6515–6529
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

cessing. They are abundant but highly redundant along successive years for a specific company or across a sector. Documents distribution across sectors is imbalanced, meaning most present sectors will have a better representation than their counterparts. Each sector has its own sociolect, independent of the phraseology of risk description which, if left alone, will be aggregated as a specific topic dimension and will spoil the risk distribution.

We propose to explore multiple types of topic models for the task of unsupervised risk extraction with regularization techniques and sector based down-sampling. In order to evaluate topic models behavior in front of these complexities, we propose two metrics, one for to quantify the coherence imbalance between sectors and another to evaluate the sensibility to the idiosyncratic sector language. Such extracted information is then validated with experts and the informativeness presented through visualization.

2. Related Works

Risk analysis is a popular task in Business and Management research (Campbell et al.; Loughran and McDonald, 2016) as such information can prevent loss due to market volatility or unexpected events (Israelsen). Risk information can be found in quantitative data like market prices where exogenous events are reflected (Fama et al., 1969), mining such information is widely used in Quantitative Finance for portfolio management and pricing. Another opportunity to find risk related information is to exploit available textual data. Specifically, researchers try to extract distributions of risks to estimate its informativeness (Singleton-Green and Hodgkinson; Campbell et al.; noa; Unknown), get a better understanding on how risk disclosures affect investor risk perception and how these correlate to markets (Koelbl et al.; Bao and Datta). On a applied perspective, the financial community showed the importance of risk disclosure for portfolio management by obtaining increased returns as compared classical factor models (Bai et al.; Lopez-Lira). Such new information, if accurately extracted, can allow to contrast market asset pricing after risk-related news, against the subjective communication of the company (Hassan et al., 2019, 2020a; Hassan et al., 2020b).

On the regulators perspective, risks distribution extraction allows for overall market study, cross-sector comparison and monitoring the evolution of risks. It is also a first step towards risk conformity verification, particularly for materiality and specificty which argumentation depends on the industry and the risk type, and attenuation such as defined in the 16th article of the Prospectus Regulation[4].

[4]ESMA guidelines on Risk Factors disclosure

To extract risk distributions, we identified three main approaches. Campbell et al. rely on a handmade dictionary and iterate on its corpus to obtain hundreds of terms related to 4 different risks. Huang and Li (2011) proposed a supervised approach with a set of 25 different risks and obtained good performances with a multi-label categorical K-nearest neighbor. Exposures defined in these documents rely on complex ontologies with interrelations between multiple types of risks, classes we could imagine are therefore overlapping depending on the granularity level at which we try to extract risk descriptions. Time is also an important factor, some risks being ignored before an exogenous shock as we've seen with Covid-19 and the pandemic risk (Loughran and McDonald; Hassan et al.). Because of those characteristics, creating a classification oriented annotated dataset or a dictionary based on pre-defined labels would have a limited interest overtime and would be restricted to the original representation granularity. In order to correct such pitfall, Bao and Datta and Israelsen proposed to use unsupervised topic modeling such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to discover risk distributions from unstructured texts. This type of modeling is still widely used in NLP applied to risk analysis in finance (Zhu et al.; Wei et al., 2019; Lopez-Lira; Bai et al.). Koelbl et al. are the first to incorporate metadata in their model with the Structural Topic Model (STM) (Roberts et al., 2016) to correct idiosyncratic terminology within industry, unfortunately they didn't evaluate to which extent this added covariate results in more reliable distributions of risks. Also, their STM is an entirely probabilistic model and does not incorporate prior knowledge such as the one found in embeddings, which are known to be helpful to obtain more coherent and diverse topics. At the best of our knowledge, there exist no open-source annotated dataset nor model that can be used for the risk analysis in the french language, increasing the need for unsupervised approaches.

Bayesian graphical probabilistic models and matrix reduction techniques were the cornerstone with the well known Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-Negative Matrix Factorization (NMF) (Pauca et al., 2004). Variants have been proposed, such as Supervised Topic Model (Mcauliffe and Blei, 2007), LabeledLDA (Ramage et al., 2009), Labeled MedLDA (Zhu et al., 2012) and Partially labeled topic models (Ramage et al., 2011; Xu et al., 2023) and have been applied to supervised tasks in specific domains (Ahmed and Xing, 2010; Aziz et al., 2022)). Other variants are meta-data based topic models which include features either to parametrize the topic distribution (Mimno and McCallum, 2012) or to add supplementary feature specific topics (Koelbl et al.). The first

one was introduced by adding feature information to condition the topic distribution depending on the observed features (Rosen-Zvi et al., 2012; Mimno and McCallum, 2012) and recently upgraded to allow unstructured features such as images as conditions (Benton and Dredze, 2018). The second was proposed by Mei et al. (2007) to separate topics from sentiments in WebBlogs and upgraded by Roberts et al. (2013). Similar systems are called *Facets Topic Models*.

Since 2017, Variational Inference has been widely used in topic modeling since it releases the constraint to build inference function for each model and let the neural network approximate the marginal distribution of latent variables. This architecture is far more flexible than probabilistic graphical topic models such as LDA and originate from the Autoencoding variational Bayes (AEVB) (Kingma and Welling, 2022; Rezende et al., 2014) and upgraded into prodLDA by (Srivastava and Sutton, 2017). In 2018 (Card et al., 2018) published the Scholar Model, which incorporates a facet topic model variant, interactions between covariates and topics, or supervision. Since Skip-Gram (Mikolov et al., 2013) and even more since BERT (Devlin et al., 2019), NLP has been in a shift towards vector-based representations, which is quite distant from the bag-of-words usual representation for topic models. Since then, authors have been incorporating embeddings into topic models (Dieng et al., 2020; Bianchi et al., 2021b,a) or into clustering models with relative success (Grootendorst, 2022; Zhang et al., 2022). Neural Topic Models are highly flexible, we advise the reader to look at the following surveys : (Abdelrazek et al., 2023; Wu et al., 2024) As Large Language Models took the lead on the last few years, we expect such models to be able to perform topic modeling. At the best of our knowledge LLMs have barely been experimented in such a way (Wang et al., 2023; Pham et al., 2023), therefore we do not include them in our study.

Although imbalance is a well known impairment for supervised and unsupervised learning, it has not been extensively studied in the topic modeling framework, especially in a multi-domain corpus. Such work on topic imbalance was initialized by Wallach et al. (2009a) who introduced an asymmetric prior on topic probabilities and demonstrated an overall increase in LDA performances. In a recent work Veselova and Vorontsov (2020) found out that *Topic Capacities* (TC), defined as the overall presence of each discovered topic in a corpus, are often close to each others in matrix factorization based probabilistic topic models. They proposed to increase the imbalance degree using a specific regularizer in the Additive Regularized Topic Model (ARTM) framework while Wu et al. (2021) proposed a causal inference approach to increase the identification of rare risks.

3. Problem Definition

We now define the task of Risk Distribution Extraction (RDE), then we detail the identified pitfalls and justify our modelisation choices. In a nutshell RDE's target is to infer the proportion of risk types per document that will be as close as possible to the true repartition of important risks presented in URDs and RDs.

Let our corpus $\boldsymbol{C}$ be composed of N pairs $\boldsymbol{x} = \{\boldsymbol{d}, v\}$ where $\boldsymbol{d_i}$ is the document i and its associated industry covariate v_i for $x_i \in C$. Each $d = \{p_i^1, p_i^2, ..., p_i^{L_i}\}$ can be separated in L_i consecutive paragraphs p. The goal of the RDE task is to approximate the unknown distribution of risks[5] $\boldsymbol{y_i} = \{y_1^i, y_2^i, ..., y_k^i\}$, where $\sum_j^k y_j^i = 1$ and k is the true number of risk types, by a model F, i.e. $F(x_i) \rightarrow \hat{y}_i$.

Considering the following modeling complexities, the redundancy of content along the years [6] that reduces the size of the corpus, the absence of existing annotations and the expected output, we consider topic modeling as the F family of models.

"Risk factors" sections of Annual reports are complex and long documents involving highly interconnected economic relations. Theoretically, a risk can be defined as a hazard with a potential for damage to an entity. Therefore it can be seen as a triplet composed of the potential event characterized as a risk, its quantitative counterparts such as the probability of occurrence, and its possible consequences (Kaplan and Garrick, 1981). As depicted, risks are the reflection of probable economic events, they inherit the complexity of economic interactions and follow a complex taxonomy. Some works have explored the possibility of creating such taxonomy (Nelson and Pritchard) resulting in 75 different risks ranging in 13 categories, one (Huang and Li, 2011) tried to annotate Annual Reports using NLP but remaining at a high level of granularity and without releasing its dataset. Separability of risks isn't an easy task and is highly dependent on the way the company creates its risk hierarchy (Appendices, Figure 6), resulting on risks being semantically close (Appendices, Figure 7).

As we dive deeper, the frontier between different classes of risks becomes shallow mainly because of causal inter-dependences between risks. Fig. 3 shows an example of *Macroeconomic environment risk* (light-blue) paragraph where the company also references "Interest-rate risk" (yellow), "Credit risk"

[5] An example of such distribution for various sectors is presented in Figure 1.

[6] Figure 2 shows the similarity of documents on consecutive years.

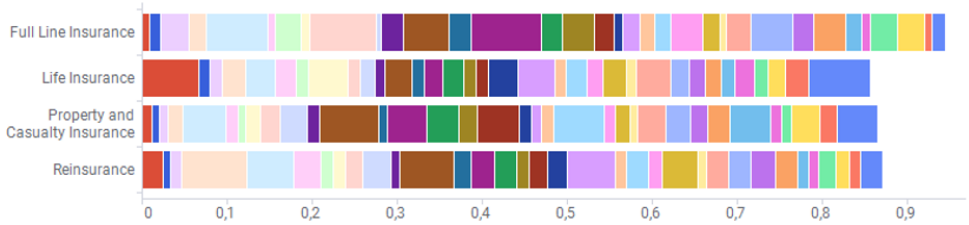

Figure 1: Financial sectors Risk Distribution example.

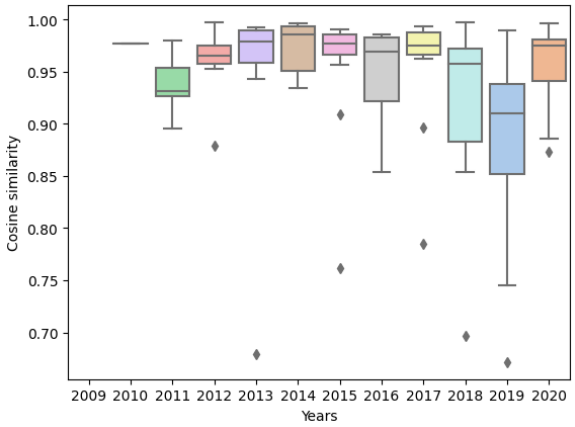

Figure 2: Average cosine simarity of documents between consecutive years.

(dark blue), "Market prices risk" (red) and "Pandemic risk" (green). The text contains 5 different risks that might also be presented elsewhere in the document.

Documents have a concentrated distributions of Risks, only a subset are presented in each one and some are far more lengthy than others. We observe around 20 risks per document with a mean and standard-deviation length of respectively around 850 and 400 tokens. Also, as we identify an important content homogeneity within sectors and a variable heterogeneity of risks between different sectors. This results in a low set of overlapping risks between economically distant sectors with more specific risks. In the Financials industry we observe around 70% of common risk factors and 30% specific to the sector. The phenomenon for Banking and Insurers is visible on Figure 4. As two sectors lie far from one another in their business aspects, their risk distributions tend to overlap less with more specific risks for each one.

Knowing these preceding corpus characteristics and that sectors are unequally distributed in our corpus (Figure 5 results in an even more contrasted risk distribution that hurt capabilities of supervised and unsupervised models. In this case, lesser represented sectors often show lower topic coherence score.

The instantiation of a risk in a given sector often involves a sector specific vocabulary such as in Insurance companies : "premium rates", "underwriting capacity" and "workers' compensation insurance". We consider this idiosyncratic terminology as out of the risk perimeter and harmful for any unsupervised topic modeling aimed at extracting distribution of risks for a given company at a given time. The domain-specific aspect of languages contained in the corpus can cause a near deterministic distribution of risks for less frequent sectors, with a low number of thematic dimensions representing the sector's specific language and its risks.

"*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 sanitary, cyclical and/or commercial reasons remains, and a lasting macroeconomic deterioration could affect the group'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.*"

Figure 3: English translation of a risk paragraph example from the 2020 URD of an insurer.

In this paper we address the problem of assessing the capability of various topic models for the extraction of companies' risks from an unbalanced multi-domain corpus which exhibits significant idiosyncratic biases.

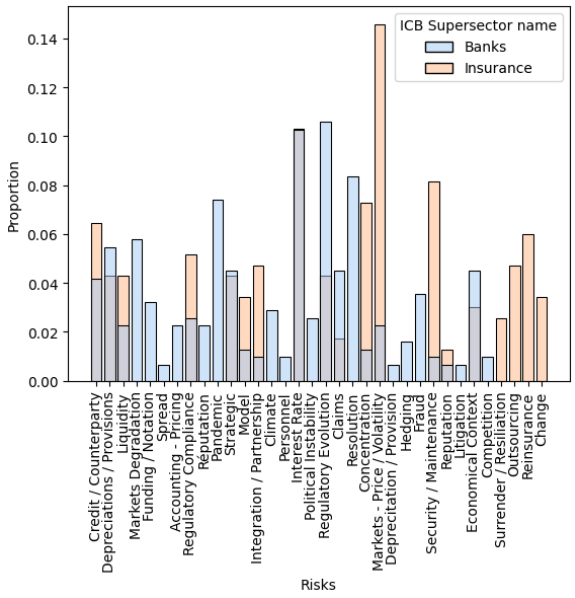

Figure 4: Bank / Insurance risk overlap.

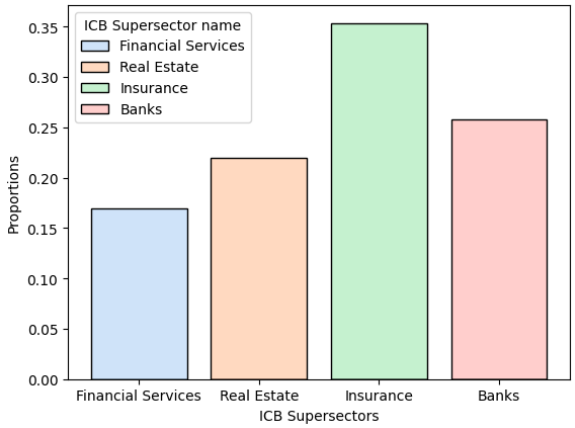

Figure 5: Financial industry documents unbalanced distribution.

4. Evaluation Methodology

4.1. Data

The dataset we used is based on a recently published corpus Masson and Paroubek (2020) and other Universal Registering Documents (URD)[7], ranging from 2013 to 2020 in the Financial industry as depicted by the ICB denomination[8]. In the paper, what we call "sectors" corresponds to "supersectors" in the ICB denomination. We manually annotated and extracted "Risk Factors" sections, when available and in a format of description close to the current legislation, which adds up to 171 documents. We then pass documents to a custom algorithm to identify paragraphs, which combines PDF information, Computer Vision and a set of rules to re-order the content. We annotated a subset of the corpus composed of two URDs per sector in the Financial industry, which sums up to 8 different documents. This dataset was annotated at the risk subsection level by assigning one of 35 different risk labels. It allows us to obtain a representative corpus to precisely evaluate extractions quality.

Most of the content of risk sections doesn't change between consecutive years, causing duplication of paragraphs : we deduplicate them with *Locality Sensitive Hashing* (Datar et al., 2004) to handle this homogeneity issue. We follow standard procedure for topic modeling text-preprocessing. We use SpaCy library for tokenization, pos-tagging and lemmatization, we also remove stop-words, verbs, non-alphanumeric and tokens appearing less than 10 times for dimensionality reduction. We use Gensim (Rehurek and Sojka, 2011) for bi-gram and tri-gram extraction with NPMI filtering. For models without background terms we remove words appearing in more than 25% of the documents as a supplementary pre-processing step. We end up with $Card(P) = 4,780$ different paragraphs with $1,335$ for Banks, $1,077$ in Insurance, 671 in Financial Services. As the cornerstone of our evaluation system in the case of unbalanced multi-domain financial corpus of french Annual Reports, we created three variants of our corpus, the first one DS_B (Balanced) with down-sampling on sectors, the second one DS_U (Unbalanced) with stratified sampling to keep the original sector imbalance and the third one C (Complete) contains the whole dataset. With them we wish to quantify how imbalance and idiosyncratic vocabulary deteriorates topic model performances and compare down-sampling to more complex models. Each dataset is then split following a 10-k fold for out-of-sample model evaluation.

4.2. Modeling

As presented previously, we believe topic modeling to be the best modeling approach with our corpus complexities. topic modeling aims at representing each input document as a distribution of $k \in K$ latent topics, each one characterized by a probabilistic distribution over words in a vocabulary V. We didn't experiment with cluster based topic modeling as we aim at approximating the Data Generating Process (DGP). Also, cluster based methods are not as flexible as neural topic models and cannot handle metadata, which is fundamental our documents DGP.

We chose at least one representative of the large classes of topic models defined previously : probabilistic, matrix and neural. As they are still widely used in Finance, we experimented with LDA (Blei

[7]Available on BDIF.
[8]Description on FTSE Russel.

et al., 2003) and NMF (Lee and Seung, 2001) for robust probabilistic and matrix-based baselines. As neural topic models recently outperformed classic models, we explored more complex and flexible approaches with the Scholar model (Card et al., 2018). This model, based on Autoencoded Variational Inference for Topic Model (AVITM) (Srivastava and Sutton, 2017) inherit the great flexibility of Variational Autoencoders (VAE) for inference, can be modified easily without having to change the inference algorithm and is able to make its computations in the embeddings space. The authors also proposed a variation of AVITM by changing the mixture to a product of experts where they allow words distributions over topics to lie in the Real space. This last change permit an easy integration of background terms, covariate specific terms and topic-covariate interactions distributions that the model can automatically learn. Topics can therefore be interprated as deviations from the background terms distribution and covariate terms distribution. Scholar is therefore a flexible model that can incorporate supervision, facets based on covariates and topics-covariates interactions.
We also explored a tailored version of Bianchi et al. (2021a) that incorporated recent contextual embeddings and accounted for covariates, yet the outcomes fell short of our expectations. During the experimentation phase, the French language lacked a robust Semantic Textual Similarity (STS) dataset suitable for fine-tuning an effective STS model, which contextualized topic models would depend on. Although recent advancements in contrastive learning have demonstrated their potential to enhance the quality of embeddings for languages without high-quality STS datasets, the improvement in performance did not justify the added a layer of complexity compared to using non-contextualized embeddings.

As we believe Scholar with covariate can improve our performances by handling idiosyncratic vocabulary per sector, we doubt it will handle the imbalance issue relative to risks. We follow Wallach et al. (2009b) and experiment different topics prior concentration parameters and topic l2 regularization coefficients for sparsity.

Based on the log-likelihood, matrix factorization based topic models such as LDA, NMF, ProdLDA and Scholar tend to exploit as much as possible each latent dimension. A latent dimension focused on a rare theme would be sub-optimal in terms of log-likelihood, or Expected Lower Bound (ELBO) in the case VAE models. These models therefore aggregate rare topics into a low set of dimensions to keep a similar capacity for each latent variable (Vorontsov and Potapenko, 2015). This problem is increased by rare sectors which tends to be concentrated into a low set of latent variables.

We end up with 5 estimated models : LDA, NMF, ProdLDA, Supervised Scholar (s-Scholar), Covariate Scholar (c-Scholar).

4.3. Metrics

We evaluate different topics quality dimensions : coherence, diversity, imbalance degree, sensibility to imbalance and sensibility to idiosyncratic sector vocabulary. As for coherence metric we rely on the widely used NPMI (Bouma, 2009) which as been demonstrated to be highly correlated with human judgement (Newman et al., 2010). As topics must be different from one another, we evaluate this dimension through diversity metric as proposed by (Nan et al., 2019).

Little work has been done on evaluating how imbalance is handled by a topic model, recent work from Veselova and Vorontsov (2020) propose to quantify the *imbalance degree* as a metric I. The *imbalance degree* is the ratio between the maximum and minimum *topic capacity* over topics and n_{p_i} is the number of tokens in the paragraph i.

$$I = \frac{n_{t_{max}}}{n_{t_{min}}},\ n_t = \sum_{\boldsymbol{i} \in \boldsymbol{C}} p(t|p_i) n_{p_i}$$

As the imbalance degree metric do not measure the coherence variation between domains we propose to evaluate imbalance sensibility of each topic model as the standard deviation of coherences evaluated on the texts of each sector. Since $NPMI \in [-1; 1]$, $\sigma_{NPMI_S} \in [0; 1]$ for which lower value means more similar coherence of topics between covariates.

$$A = \sqrt{\frac{1}{|M|} \sum_{m=1}^{|M|} (NPMI_m - \overline{NPMI})^2}$$

To evaluate the sensibility to idiosyncratic vocabulary we quantify the topic concentration per covariate as the mean of Kullback-Leibler Divergences between each average distribution of topics per covariate and a discrete uniform distribution. It allows us to quantify, in terms of bits of information, how concentrated our topics distributions depending on each covariate. Great values demonstrate highly concentrated risks per sector and low overlap, which means our latent variables are not shared between sectors and are too sensible to sector semantics.

$$ToCo = \frac{1}{|M|} \sum_{m=1}^{|M|} KL(P(Z|m)||U(|M|))$$

5. Experimentation

In this experimentation to evaluate topic models behavior in the case of multi-domain language, domain and topics imbalance, we split the problem in three questions : (1) "*In terms of domain imbalance and domain specific vocabulary, how are classic topic models impacted ?*", (2) "*How helpful*

are down-sampling, supervision and covariates integration to topic models for handling these pitfalls ?" and (3) "*When taking topic quality into account, should we prefer down-sampling, supervision or covariates ?*"

We trained LDA and NMF models with Gensim. Scholar (Card et al., 2018) and ProdLDA (Srivastava and Sutton, 2017) models were trained for 450 epochs, with a learning rate of 0.02, batch size of 200 and embeddings from a skip-gram model with $d = 300$ trained on our preprocessed and deduplicated corpus. As hyperparameters tuning and regularization, we put a low symmetric alpha parameter on the topic prior $\alpha = 0.3$ to force the concentration of topic probabilities and a low l1 regularization $l1 = 0.1$ on covariates words representation. For the TopicPrior regularizer we use a Dirichlet prior for $beta$ with concentration parameter $\gamma = 0.3$ and a $\tau = 0.3$. We chose $k = 50$ as the number of topics, which was in line was what was expected from the data. We then evaluated all models and dataset variants on our metrics, see Figure 1.

Results of our experimentation are presented in Table 1. Classic topic models, such as LDA and NMF, shows non null CoIm metric when trained on an unbalanced corpus which means NPMI vary significantly on different sectors. Interestingly, this metric decreases with more data as we can see comparing these models between the DS_U and C datasets. About ToCo we find high sensibility to multi-domain language for both models no matter the training dataset. They seem to create topics that are very sector specific but quite general in the sector, as the low diversity suggests particularly for NMF. prodLDA and scholar models are also sensible to imbalance but not much to domain-vocabulary for which the metric is 5 to 9 times lower.

Down-sampling offers the best performance on coherence imbalance, but in DS_U and C datasets c-Scholar also helps to reduce these biases. Down-sampling has barely no effect on domain vocabulary sensibility but adding more data, even unbalanced, is useful for c-Scholar which grants the lowest ToCo metric with 0.19. Covariates as word distributions helps to get an overlapping distribution of themes between sectors but has little positive impact on imbalance reduction.

As we've seen the interest in using covariate topic models to get less biased topics and lower the imbalance, incorporating coherence metrics and diversity can help us to decide whether down-sampling or using more complex models is the appropriate solution. Coherence metrics are significantly lower in the down-sampled dataset as for diversity and the best performances are achieved when training on the complete dataset. NPMI is hardly comparable between the covariate model and all the others, some important words being moved from the topics words distributions to the sector specific word distribution, but neural topic models show better overall coherence. Diversity is maximised for covariate Scholar with a gain of at least 10 points in each dataset. Covariate models on a complete and asymmetric dataset tend to offer a better diversity of topics, a lower imbalance in coherence and a great capacity to handle the multiplicity of domains without collapsing into sector specific topics, some examples are shown in Table 2.

6. Results analysis

Based on the detailed evaluation we went through in the previous section, we selected the c-Scholar model and trained it on the overall corpus for an in depth analysis of risk factors along in the Financial sector from 2013 to 2020. Some conclusions of the interest of such system for the French Financial Market Supervisor has been published in a special note[9].

Informations about risk distributions extracted after this study were incorporated in a custom interface for monitoring by the expert teams. The interface was built around augmenting analysis capabilities of experts and information discovery, various dashboards descriptions are presented in 3. Many of the representations can be found in the special note published on AMF's website.

The results of the model also make it possible to explore the change in risks over time. The figure 9 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.

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 (Figure 10).

Following the previous example, we can see on Figure 8 that the selected document presents significantly more "Climate Risks" than its sectors counterpart. It can be explained by the fact that this year they moved their "Climate Risk" sections from its Declaration of Extra-Financial Performance (DEFP) to its "Risk Factors" section.

[9] "Automated risk factor analysis published by listed companies: a use case of NLP for the AMF" - Link

		NPMI	Diversity	CoIm	ToCo	Purity	I
DS_U	NMF	-0.019	0.422	0.034	0.178	0.399	5.035
	LDA	0.025	0.645	0.030	0.279	0.437	12.399
	prodLDA	0.017	0.685	0.031	0.036	0.462	3.602
	s-Scholar	**0.030**	0.628	0.031	0.034	**0.543**	3.683
	c-Scholar	-0.099	**0.797**	**0.025**	**0.026**	0.463	3.877
C	NMF	0.005	0.437	0.029	0.204	0.416	5.532
	LDA	0.016	0.660	**0.020**	0.293	0.451	12.096
	prodLDA	0.024	0.785	0.043	0.036	**0.466**	3.202
	s-Scholar	**0.031**	0.760	0.040	0.040	0.452	3.089
	c-Scholar	-0.053	**0.894**	0.037	**0.019**	0.465	3.076
DS_B	NMF	-0.042	0.414	0.012	0.174	0.393	5.981
	LDA	0.014	0.597	0.010	0.263	0.438	13.880
	prodLDA	0.018	0.700	0.005	0.043	**0.453**	3.433
	s-Scholar	**0.022**	0.685	**0.003**	0.053	0.437	3.594
	c-Scholar	-0.084	**0.802**	0.006	**0.034**	0.442	3.652

Table 1: Evaluation of topic models

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

Table 2: Keywords examples using c-Scholar with 50 topics.

Interface tab	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.

Table 3: List of analysis criteria made possible by the display of results.

7. Conclusion

Using a specific dataset of french corporate filings in the financial industry with various pitfalls (imbalance, idiosyncratic sector vocabulary, redundancy, ...), we explored different topic models, regularizations and dataset construction methods and evaluated them in terms of coherence, diversity, sensibility to imbalance and to subdomain language. We found that down-sampling is currently the best way to correct imbalance of sector sizes. Otherwise, covariate model based on SCHOLAR architecture and trained on the complete dataset offers the best performances on comparable metrics and particularly for diversity and resilience to idiosyncratic vocabulary. We also presentend parts of an interface for in-depth investigation and knowledge discovery whose ergonomy is designed with supervision analysts in mind. The interface allows to identify new information that is difficult to spot manually and redirect attention towards documents that are outliers in terms of risk content and temporal risk distribution evolution.

8. Ethical considerations and limitations

The ethical risks associated with the work presented herein are in our opinion quasi-inexistant since our research uses only algorithm which do not require external datasets in complement to the input document. Nevertheless the experiments rely on a corpus we collected, so there is always the possibility of an underlying selection bias, but we took great care to perform the widest and most homogeneous filtering from the sources. Because the document we work with are public, freely available from a AMF website [10] and required by regulations, their form and content are assumed to be fully compliant with GDPR regulations and thus devoid of any risk of infringement on privacy.
This work is a research and such system should not be considered as reliable for investment decisions.

9. Acknowledgements

This research is funded by a collaboration between the French Financial Market Authority (AMF) and the LISN laboratory from CNRS associated with Paris-Saclay University.

[10] BDIF website Link

10. Bibliographical References

The corporate risk factor disclosure landscape.

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Amr Ahmed and Eric Xing. 2010. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1140–1150, Cambridge, MA. Association for Computational Linguistics.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Saqib Aziz, Michael Dowling, Helmi Hammami, and Anke Piepenbrink. 2022. Machine learning in finance: A topic modeling approach. *European Financial Management*, 28(3):744–770.

Hebatallah Badawy and Adel Nematallah Ibrahim. 2016. Is the readability of corporate textual disclosures measurable? *SSRN Electronic Journal*.

John (Jianqiu) Bai, Priya Garg, Sarah Shaikh, and Chi Wan. Overlapping narrative risk disclosures and return predictability. (ID 3821163).

Scott R. Baker, Nicholas Bloom, and Steven J. Davis. 2016. Measuring economic policy uncertainty*. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Ravi Bansal, Dana Kiku, and Marcelo Ochoa. 2016. Price of long-run temperature shifts in capital markets. Working Paper 22529, National Bureau of Economic Research.

Yang Bao and Anindya Datta. Simultaneously discovering and quantifying risk types from textual risk disclosures.

Adrian Benton and Mark Dredze. 2018. Using author embeddings to improve tweet stance classification. In *EMNLP Workshop on Noisy User-generated Text (W-NUT)*, pages 184–194.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022.

Patrick Bolton and Marcin Kacperczyk. 2021. Do investors care about carbon risk? *Journal of Financial Economics*, 142(2):517–549.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

John L. Campbell, Hsinchun Chen, Dan S. Dhaliwal, Hsin-min Lu, and Logan B. Steele. The information content of mandatory risk factor disclosures in corporate filings. 19(1):396–455.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural Models for Documents with Metadata. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040. ArXiv: 1705.09296.

Jaiho Chung, Hyungseok Kim, Woojin Kim, and Yong Keun Yoo. 2012. Effects of disclosure quality on market mispricing: Evidence from derivative-related loss announcements. *Journal of Business Finance & Accounting*, 39(7-8):936–959.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, SCG '04, pages 253–262, New York, NY, USA. Association for Computing Machinery.

João Antônio Salvador de Souza, Jean Carlo Rissatti, Suliani Rover, and José Alonso Borba. 2019. The linguistic complexities of narrative accounting disclosure on financial statements: An analysis based on readability characteristics. *Research in International Business and Finance*, 48:59–74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1041–1048, Bellevue, Washington, USA. Omnipress.

Renato Faccini, Rastin Matin, and George Skiadopoulos. 2021. Dissecting climate risks: Are they reflected in stock prices? *SSRN Electronic Journal*.

Eugene F. Fama, Lawrence Fisher, Michael C. Jensen, and Richard W. Roll. 1969. The adjustment of stock prices to new information. *SSRN Electronic Journal*.

Evgenii Gorbatikov, Laurence van Lent, Narayan Y. Naik, Varun Sharma, and Ahmed Tahoun. 2019. Is firm-level political exposure priced? *SSRN Electronic Journal*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Sanford J. Grossman and Joseph E. Stiglitz. 1980. On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408.

Tarek A. Hassan, Stephan Hollander, Laurence van Lent, Markus Schwedeler, and Ahmed Tahoun. Firm-level exposure to epidemic diseases: COVID-19, SARS, and h1n1. (ID 3566530).

Tarek A Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2019. Firm-Level Political Risk: Measurement and Effects*. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Tarek Alexander Hassan, Stephan Hollander, Laurence van Lent, Markus Schwedeler, and Ahmed Tahoun. 2020a. Firm-Level Exposure to Epidemic Diseases: COVID-19, SARS, and H1N1. Working Paper 26971, National Bureau of Economic Research.

Tarek Alexander Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2020b. The

Global Impact of Brexit Uncertainty. Working Paper 26609, National Bureau of Economic Research.

Harrison Hong, Frank Weikai Li, and Jiangmin Xu. 2019. Climate risks and market efficiency. *Journal of Econometrics*, 208(1):265–281. Special Issue on Financial Engineering and Risk Management.

Ke-Wei Huang and Zhuolun Li. 2011. A multilabel text classification algorithm for labeling risk factors in sec form 10-k. *ACM Trans. Management Inf. Syst.*, 2:18.

Ryan D. Israelsen. Tell it like it is: Disclosed risks and factor portfolios. (ID 2504522).

Stanley Kaplan and B. John Garrick. 1981. On The Quantitative Definition of Risk. *Risk Analysis*, 1(1):11–27.

Diederik P Kingma and Max Welling. 2022. Auto-encoding variational bayes.

Marina Koelbl, Ralf Laschinger, Bertram I. Steininger, and Wolfgang Schäfers. Revealing the risk perception of investors using machine learning. (ID 3686492).

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–53, Gothenburg, Sweden. Association for Computational Linguistics.

Daniel Lee and H. Sebastian Seung. 2001. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Andrea Liesen, Frank Figge, Andreas Hoepner, and Dennis M. Patten. 2016. Climate change and asset prices: Are corporate carbon disclosure and performance priced appropriately? *Journal of Business Finance & Accounting*, 44(1–2):35–62.

Alejandro Lopez-Lira. Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. (ID 3313663).

Tim Loughran and Bill McDonald. Management disclosure of risk factors and COVID-19. (ID 3575157).

Tim Loughran and Bill McDonald. 2016. Textual Analysis in Accounting and Finance: A Survey. SSRN Scholarly Paper ID 2504147, Social Science Research Network, Rochester, NY.

Ella Mae Matsumura, Rachna Prakash, and Sandra C. Vera-Muñoz. 2014. Firm-value effects of carbon emissions and carbon disclosures. *The Accounting Review*, 89(2):695–724.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, page 171, Banff, Alberta, Canada. ACM Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [cs].

David Mimno and Andrew McCallum. 2012. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv:1206.3278 [cs, stat]*. ArXiv: 1206.3278.

Irene Monasterolo and Luca de Angelis. 2020. Blind to carbon risk? an analysis of stock market reaction to the paris agreement. *Ecological Economics*, 170:106571.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy. Association for Computational Linguistics.

Karen K. Nelson and Adam C. Pritchard. Litigation risk and voluntary disclosure: The use of meaningful cautionary language. (ID 998590).

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

V. Paul Pauca, Farial Shahnaz, Michael W. Berry, and Robert J. Plemmons. 2004. Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore. Association for Computational Linguistics.

Daniel Ramage, Christopher D. Manning, and Susan Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*. ACM Press.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models.

M. Roberts, B. Stewart, D. Tingley, and E. Airoldi. 2013. The structural topic model and applied social science. *Neural Information Processing Society*.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2012. The author-topic model for authors and documents. *arXiv:1207.4169 [cs, stat]*. ArXiv: 1207.4169.

Zacharias Sautner, Laurence van Lent, Grigory Vilkov, and Ruishen Zhang. 2021. Firm-level Climate Change Exposure. SSRN Scholarly Paper ID 3642508, Social Science Research Network, Rochester, NY.

Brian Singleton-Green and Robert Hodgkinson. Reporting business risks: Meeting expectations. (ID 2330378).

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models.

Nassim Nicholas Taleb. 2007. *The Black Swan: The Impact of the Highly Improbable*. Random House Group.

Unknown. Risk identification: What is in the 10-k?

Eugeniia Veselova and Konstantin Vorontsov. 2020. Topic balancing with additive regularization of topic models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, page 59–65. Association for Computational Linguistics.

Konstantin Vorontsov and Anna Potapenko. 2015. Additive regularization of topic models. *Machine Learning*, 101(1):303–323.

Hanna Wallach, David Mimno, and Andrew McCallum. 2009a. Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Hanna Wallach, David Mimno, and Andrew McCallum. 2009b. Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. 2023. Prompting large language models for topic modeling.

Lu Wei, Guowen Li, Jianping Li, and Xiaoqian Zhu. Bank risk aggregation with forward-looking textual risk disclosures. 50:101016.

Lu Wei, Guowen Li, Xiaoqian Zhu, Xiaolei Sun, and Jianping Li. 2019. Developing a hierarchical system for energy corporate risk factors based on textual risk disclosures. *Energy Economics*, 80:452–460.

Xiaobao Wu, Chunping Li, and Yishu Miao. 2021. Discovering topics in long-tailed corpora with causal intervention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 175–185, Online. Association for Computational Linguistics.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

Weijie Xu, Xiaoyu Jiang, Srinivasan H. Sengamedu, Francis Iannacci, and Jinjin Zhao. 2023. vONTSS: vMF based semi-supervised neural topic modeling with optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4433–4457. ArXiv:2307.01226 [cs, math].

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2012. Medlda: Maximum margin supervised topic models. *J. Mach. Learn. Res.*, 13(1):2237–2278.

Xiaodi Zhu, Steve Y. Yang, and Somayeh Moazeni. Firm risk identification through topic analysis of textual financial disclosures. (ID 2820313).

11. Language Resource References

Masson, Corentin and Paroubek, Patrick. 2020. *NLP Analytics in Finance with DoRe: A French 250M Tokens Corpus of Corporate Annual Reports*. European Language Resources Association.

12. Appendices

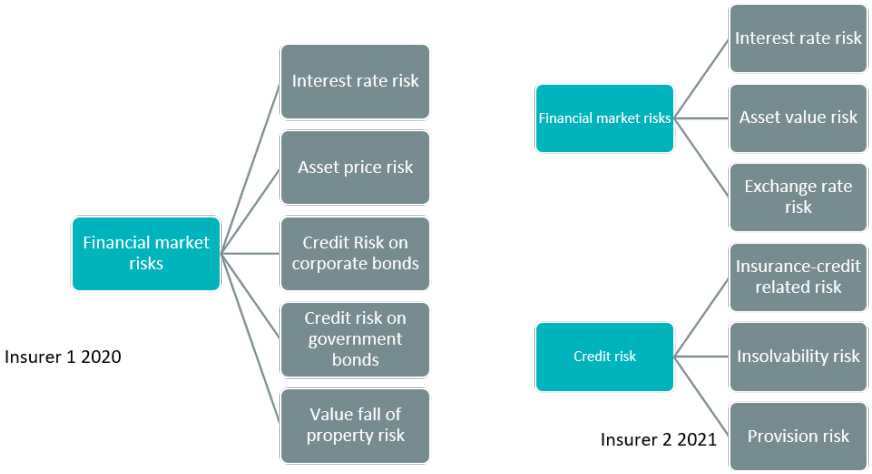

Figure 6: Comparison between two presentations of financial risks.

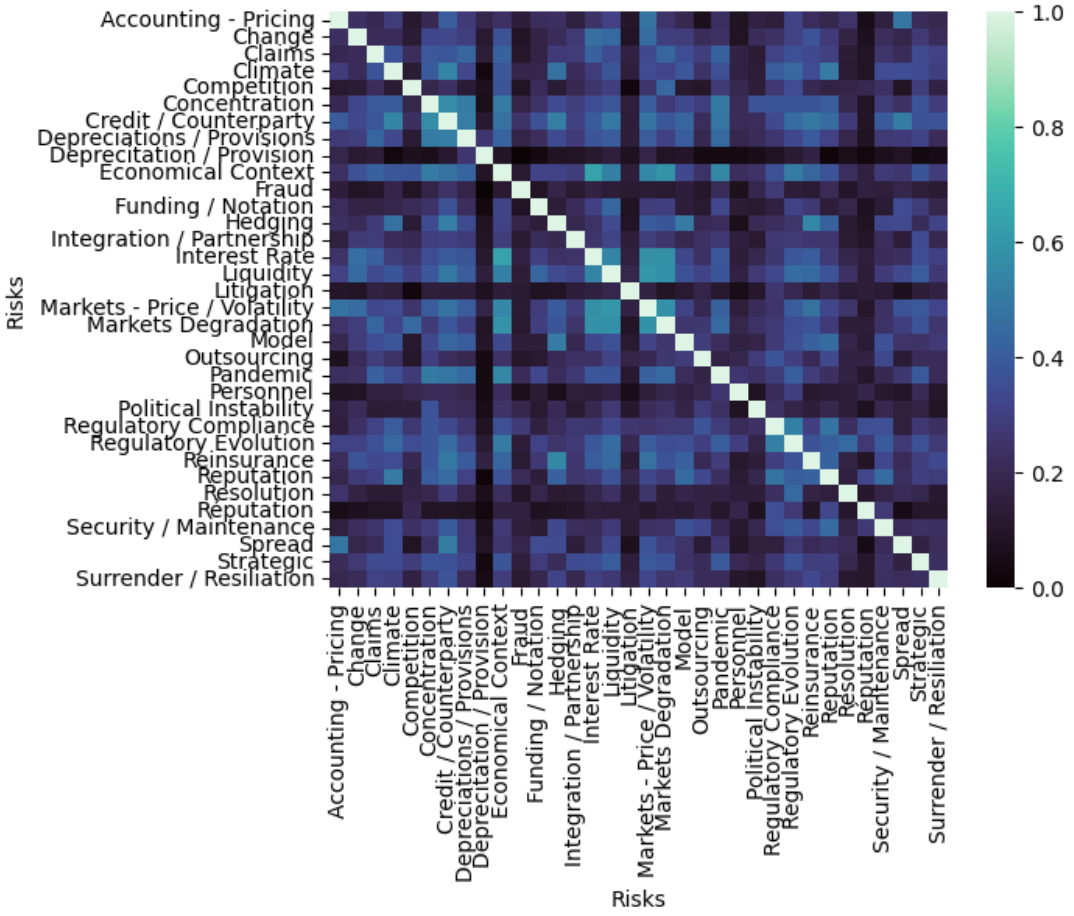

Figure 7: Mean cosine similarity between pairs of types of risks.

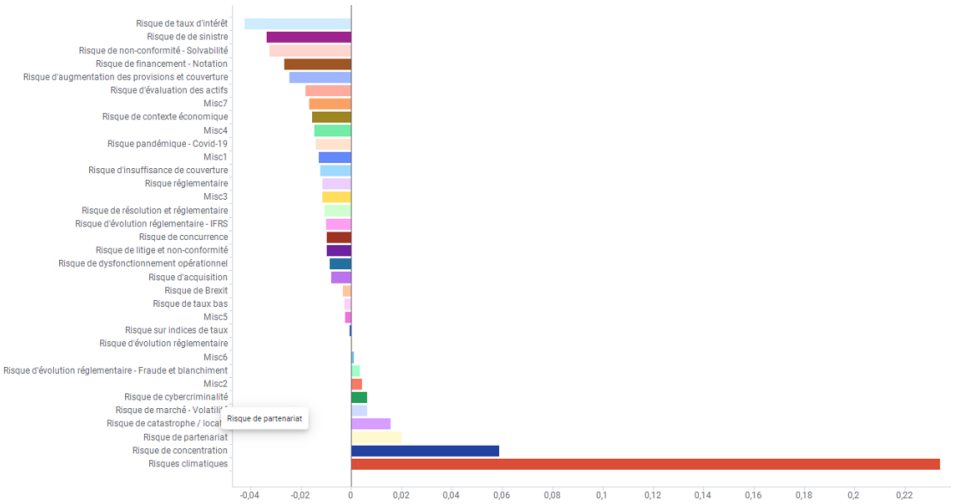

Figure 8: Risk distribution divergence between an URD and its sector profile.

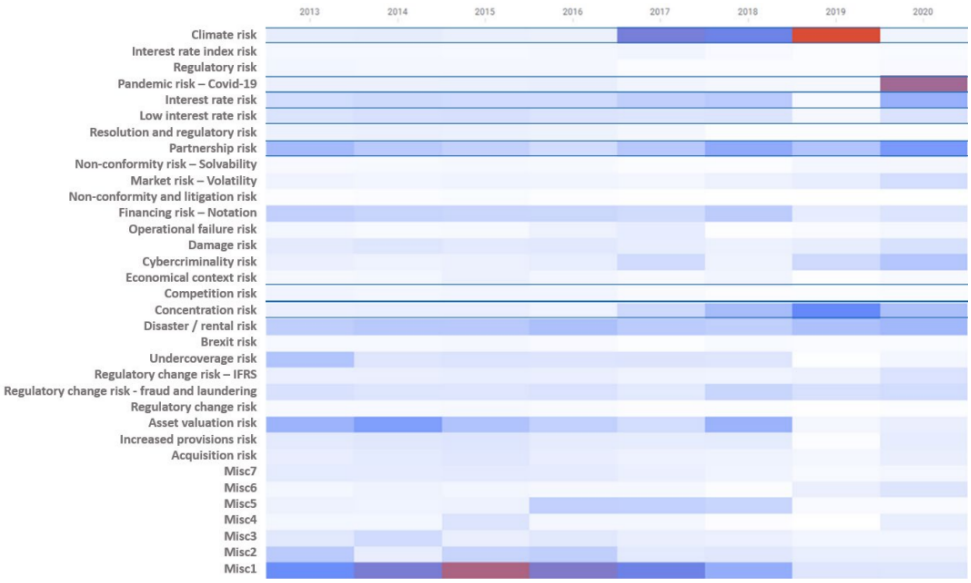

Figure 9: Risk distribution evolution between acounting years 2013 and 2020.

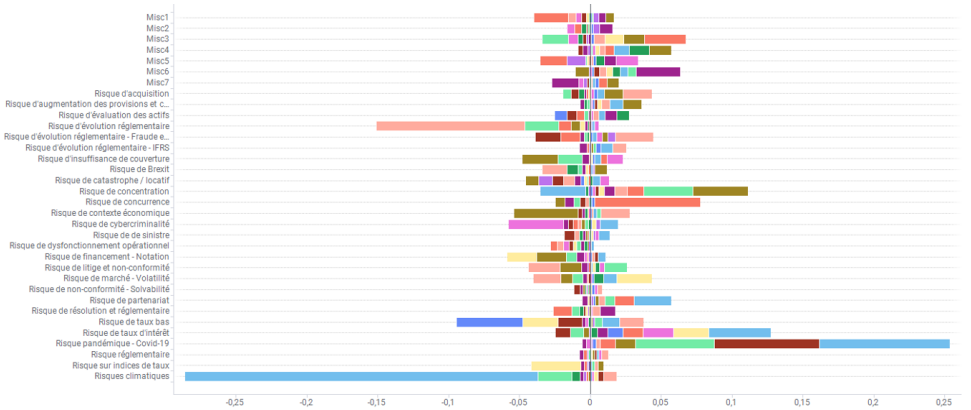

Figure 10: Variation of the risk factors distribution reported between ccounting year 2019 and 2020.