nexialog CONSULTING

NEXIASEARCH

Analyse ESG assistée par NLP : cas pratique sur données et modèles open-source

Anas ZILALI

THINK SMART ACT DIFFERENT

TABLE DES MATIÈRES

INTRODUCTION

L'ESG est un terme international utilisé par la communauté financière pour désigner les critères Environnementaux, Sociaux et de Gouvernance (ESG) qui constituent généralement les trois piliers de l'analyse extra-financière. Ils permettent d'évaluer l'exercice de la responsabilité d'une organisation et de ses parties prenantes (salariés, partenaires, sous-traitants et clients) vis-à-vis de ses décisions et de ses activités impactant l'environnement, la société et la gouvernance.

L'évaluation d'une entreprise est essentielle pour les décisions d'investissement, non seulement pour prendre en compte les risques durables susceptibles d'affaiblir sa solidité financière, mais aussi pour évaluer l'impact du produit sur des questions clés qui représentent un risque systémique pour la société, comme le changement climatique, la fraude, la corruption, et la cohésion sociale. Il est ainsi nécessaire, dans une stratégie inscrite dans le développement durable et l'investissement responsable, de relier la performance financière d'une entreprise à son impact environnemental et social.

Pour réaliser cette évaluation, il est primordial d'avoir accès aux données ESG des entreprises, ce qui est souvent complexe.

Cette note présente une illustration de l'application du traitement du langage naturel (NLP - Natural Language Processing) dans le cadre de l'analyse ESG.

La Section I met en évidence les défis liés à la collecte des données ESG et présente plusieurs sources de données en libre accès (actualités, publications etc.) pouvant être utilisées pour tirer profit du NLP dans ce cadre. La Section II aborde le topic modeling et la classification du texte, en expliquant comment ces deux traitements sont utilisés dans l'analyse ESG. Elle présente ensuite l'analyse de sentiment, en réalisant une revue des modèles basés sur BERT (un modèle de compréhension de la langue avec une architecture Transformer [7]) et en mettant l'accent sur les modèles adaptés à la finance et aux critères ESG.

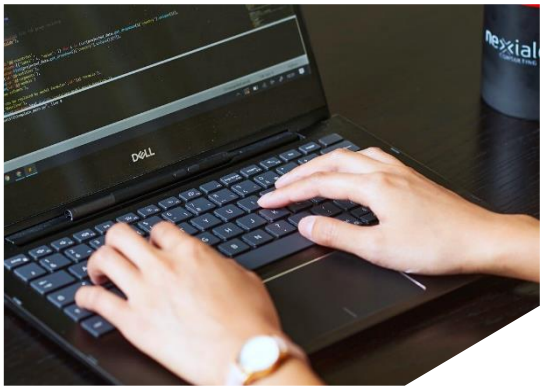
Afin d'illustrer le potentiel du NLP pour l'analyse ESG, la Section III présente un cas d'usage sur des données collectées à partir de l'API du New York Times [3]. Ce dernier est l'un des journaux les plus influents du monde, il couvre l'actualité américaine et internationale, la politique, l'économie, la finance, les sciences, les technologies et bien d'autres sujets. Bien que le journal ne soit pas spécialisé dans la finance et l'ESG, il propose (contrairement aux autres journaux) d'exploiter ces données dans un cadre non commercial [1] avec une API qui simplifie l'accès aux métadonnées des publications. En plus, la plupart des algorithmes NLP disponibles en open source sont exclusivement adaptés à l'anglais, ce qui limite leur utilisation sur des journaux français. La Section III conclut sur les risques et les limites associés à l'utilisation de ces modèles et l'importance des jeux de données utilisés lors de l'inférence.

I. Les défis de la collecte des données ESG et le potentiel du NLP pour renforcer ESG

Les défis de la collecte des données ESG

Il est actuellement difficile d'obtenir des données brutes ESG. D'abord, certaines entreprises ne publient pas de rapports ESG, ou ne fournissent pas d'informations complètes et détaillées sur leur performance ESG.

Les informations ESG peuvent également être dispersées dans différents rapports ou sur différents sites web de l'entreprise, ce qui rend la collecte de ces informations plus difficile. En plus, les entreprises peuvent être réticentes à divulguer des informations sensibles qui pourraient les exposer à des risques juridiques pouvant nuire à leurs réputations. Enfin, malgré les efforts des pays à structurer les Reportings ESG [29], il n'existe pas encore de normes universelles en vigueur pour la collecte, la présentation et la communication des données ESG, ce qui rend difficile la comparaison des performances ESG d'une entreprise à une autre.



Le potentiel du NLP

Il existe néanmoins des sources de données textuelles extra-financières qui sont sous-utilisées et qui pourraient, si elles sont correctement analysées, renforcer l'évaluation des critères ESG.

L'intelligence artificielle, et plus particulièrement le traitement du langage naturel (NLP), pourrait permettre aux investisseurs d'exploiter ces données extra-financières de manière plus efficace.

Le traitement du langage naturel (NLP) est une branche de l'intelligence artificielle qui s'attache à donner la capacité aux machines d'interpréter, générer ou traduire le langage humain tel qu'il est écrit et/ou parlé. Le NLP est utilisé pour identifier, extraire et analyser des informations à partir de données textuelles non structurées telles que les rapports annuels, les rapports de développement durable des entreprises et les articles d'actualité.

Le NLP peut augmenter considérablement la portée de l'analyse ESG sur les données extra-financières. Ces données permettent également de confronter différentes sources d'information afin d'assurer la conformité des renseignements transmis dans les rapports de gestion des entreprises. C'est en outre un moyen de certifier et valider les informations publiées dans ces rapports.

Le NLP appliqué sur ces données est donc particulièrement utile sur le périmètre des marchés émergents où l'accès à de la donnée ESG peut être limité.

En plus, ces données extra-financières pourraient aider à lutter contre la fatigue du reporting ESG (un épuisement ou une lassitude de l'entreprise vis-à-vis de la communication des informations liées à aux critères ESG [28]).

Les techniques NLP peuvent aider à identifier de manière temporelle les risques environnementaux, sociaux et de gouvernance et à mettre en lumière les aspects clés du texte et de son analyse sentimentale. Cela contribue à la détection précoce des risques ESG, à l'identification des mesures d'atténuation des risques ESG, des stratégies d'engagement et de transition ainsi que la surveillance et le suivi du portefeuille d'investissement.

Plusieurs solutions de marché exploitent les techniques NLP pour extraire les informations ESG à partir des rapports de gestion, des actualités en lien avec l'activité de l'entreprise. Les offres basées sur le NLP vont de la surveillance des risques en temps réel à l'extraction d'indicateurs liés aux objectifs de développement durable et aux scores ESG. Les fournisseurs de service de scoring ESG basés sur le NLP incluent TruValue Labs (acquis par FactSet [20] en 2020), Datamaran [21], Entis [22], MSCI [23] et Sesamm [24].

II. Les techniques du NLP

Les tâches courantes du NLP sont l'analyse de sentiment, la classification de texte, la reconnaissance d'entités nommées (NER - Named-entity recognition) et le topic modeling (la modélisation thématique). Cette Section présente un aperçu sur le Topic modeling, la classification de texte et l'analyse de sentiment

Topic modeling et classification de texte

Le topic modeling (modélisation thématique) est une technique de NLP qui permet de découvrir les thèmes sous-jacents dans un ensemble de documents.

Il s'agit d'une approche non supervisée qui utilise des algorithmes statistiques pour identifier les sujets les plus courants dans un corpus de textes.

Les algorithmes les plus couramment utilisés pour le topic modeling incluent Latent Dirichlet Allocation (LDA) [4], Probabilistic Latent Semantic Analysis (PLSA) [5] et Hierarchical Dirichlet Process (HDP) [6].

Chacun de ces algorithmes utilise des techniques différentes pour extraire les thèmes dans un corpus de textes, telles que l'analyse de la fréquence des mots, la reconnaissance des relations entre les mots et l'analyse des contextes.

La principale différence entre la classification de texte et le topic modeling est que la classification est une méthode d'apprentissage supervisé qui consiste à affecter une étiquette prédéfinie à chaque document en utilisant un modèle entraîné antérieurement sur cette tâche, tandis que le topic modeling consiste à regrouper les données similaires sous forme de cluster.

Le topic modeling ainsi que la classification de texte peuvent être utilisés pour détecter les passages d'un document ayant un lien thématique avec les critères E, S et G, pour analyser la couverture médiatique des sujets ESG et l'évaluation de la qualité de l'information ESG. Pour obtenir de bons résultats, il est souvent nécessaire de prétraiter et de nettoyer les données pour éliminer les bruits et les distractions inutiles (par exemple, cela peut impliquer l'élimination des mots vides tels que "le" ou "de" qui n'apportent pas beaucoup de sens aux données), et de choisir un algorithme du topic modeling ou de classification adapté aux données et aux objectifs (si les classes ne sont pas prédéfinies un algorithme de topic modeling est plus pertinent dans ce cas).

Analyse de sentiment

L'analyse de sentiment est une technique qui consiste à classer un texte comme porteur d'un sens positif ou négatif grâce à l'utilisation du traitement du langage naturel (NLP). Dans le domaine de la finance et de l'ESG, les capacités de compréhension du langage naturel (NLU - Natural Language Understanding) sont nécessaires pour automatiser les tâches d'analyse qui peuvent être ensuite utilisées pour construire des scores ESG.

BERT et ses variants

BERT (Bidirectional Encoder Representations from Transformers) :

BERT [7] est un modèle de langage développé par le département 'AI Language' de Google en 2019. BERT est capable d'apprendre des représentations de mots (sous forme de vecteurs dans un espace vectoriel d'une taille fixe) à partir de grands volumes de textes non annotés. Contrairement aux approches précédentes, les représentations BERT sont hautement contextuelles en raison de leur formulation profonde et bidirectionnelle (c'est-à-dire de gauche à droite et de droite à gauche). BERT a été pré-entraîné avec deux objectifs :

- Le modèle de langage masqué (MLM - Masked Language Modeling) pour entraîner le modèle à comprendre une représentation bidirectionnelle de la phrase. 15% des mots sont masqués de manière aléatoire dans l'entrée et le modèle est entraîné à prédire les mots masqués.
- Prédiction de la phrase suivante (NSP – Next Sentence Prediction), où le modèle est entraîné pour prédire si les deux phrases d'entrée se suivent naturellement.

Ces deux objectifs permettent au modèle d'apprendre la représentation de la langue pour extraire des caractéristiques utiles pour d'autres tâches NLP.

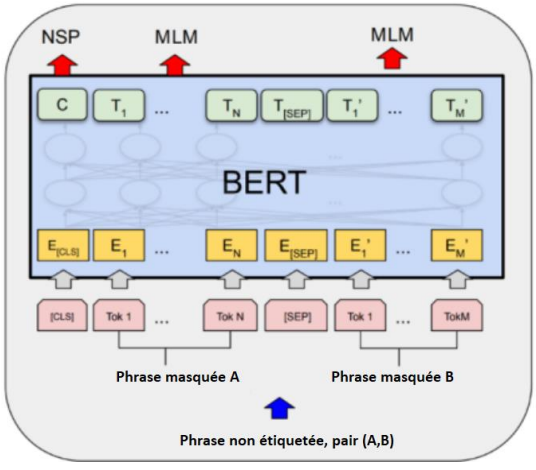
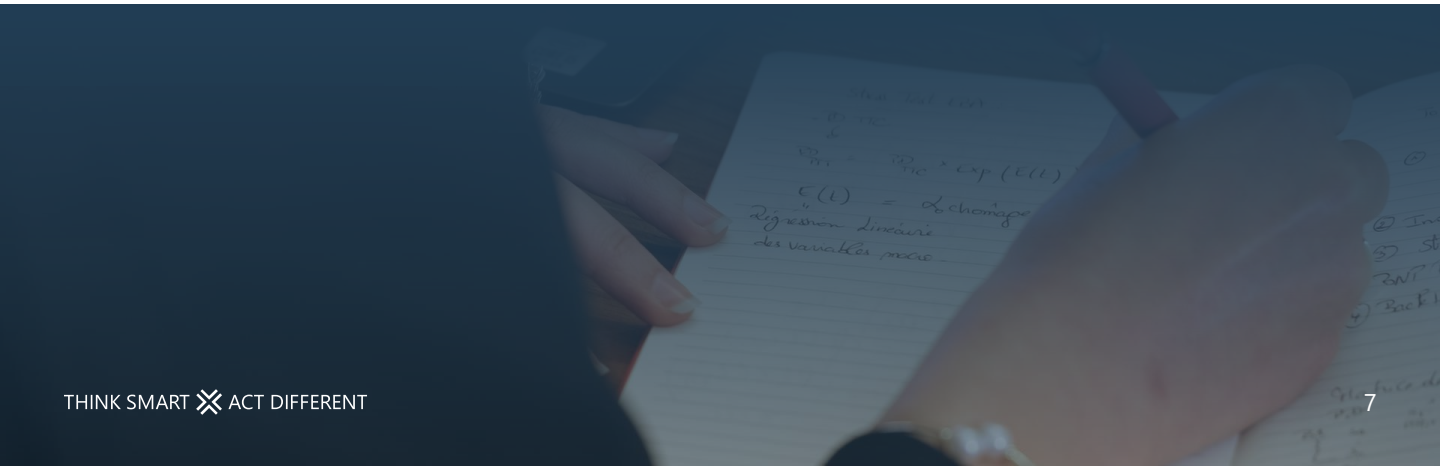


Figure 1 : Procédure globale de pré-entraînement de BERT [7]



En plus, BERT exploite une architecture d'encodeur-décodeur (transformer) qui utilise des mécanismes d'attention pour transmettre au décodeur une image plus complète de l'ensemble de la séquence en une seule fois plutôt que de manière séquentielle. Plusieurs variantes du modèle BERT ont été développées par la suite pour améliorer les performances du NLP. RoBERTa et DistilRoBERT constituent deux des améliorations les plus représentatives.

RoBERTa [8] (Robustly Optimized BERT Pre-training Approach) est une version améliorée du modèle BERT. Les principales différences entre RoBERTa et BERT se situent dans la taille du corpus et les techniques de pré-entraînement utilisées. RoBERTa a été entraîné sur un corpus de texte beaucoup plus grand et plus diversifié, ce qui lui permet d'avoir une compréhension plus profonde du contexte et des relations sémantiques entre les mots. Contrairement à BERT, RoBERTa ne repose pas sur du NSP et change dynamiquement les masques dans l'étape du MLM.

DistilRoBERTa [9,10] est une version distillée du modèle RoBERTa qui conserve ses capacités de représentation du langage tout en étant plus rapide et plus léger. Cela en fait un choix attractif pour les applications NLP ayant des ressources de calcul limitées.

Adaptation de domaine

Plusieurs études dans différents domaines ont démontré que les adaptations de BERT à un domaine spécifique sont plus performantes que la version BERT générique [7]. Cela permet au modèle de s'adapter aux particularités du langage et aux conventions spécifiques à ce domaine, ce qui peut améliorer ses performances sur des tâches liées à ce domaine. Les adaptations les plus connues sont BioBERT dans le domaine biomédical et SciBERT dans le domaine scientifique [11, 12].

Le tableau suivant présente une revue des modèles basés sur BERT [7] et qui sont utilisés dans le domaine de la Finance et des ESG pour effectuer différentes tâches (plus d'informations dans l'Annexe).

Modèle	Tâches	Démarche	Open-source
FinEAS [13]	Analyse de sentiment	Fine-tuning de sentence-BERT sur des données d'actualités financière annotées fournies par l'entreprise RavenPack [25] (pas de pré-entraînement)	Oui
DistilRoberta Finetuned [14]	Analyse de sentiment	Fine tuning de distillRoberta sur la base de données Financial Phrasebank (pas de pré-entraînement)	Oui
FinancialBERT [15]	Analyse de sentiment	Pré-entraînement de BERT sur un large corpus de textes financiers (Bloomberg News, TRC2-financial ...) et fine-tuning sur Financial Phrasebank.	Oui
SEC-BERT Finetuned [16]	Analyse de sentiment	Pré-entraînement de BERT sur des rapports 10-K de "U.S. Securities and Exchange Commission (SEC)", et fine-tuning sur Financial Phrasebank + une base de Covid19 sur Kaggle.	Oui
FinBERT [17]	Analyse de sentiment + Classification ESG	Pré-entraînement de BERT sur un corpus de communication financière (rapports d'entreprise 10-K & 10-Q, transcriptions d'appels sur les revenus ...) puis fine-tuning sur les tâches en aval.	Oui
ESGBERT [18]	Classification changement / pas de changement + classification positive / négative	Pré-entraînement de BERT sur des données de "Knowledge Hub" du projet "Accounting for Sustainability" puis fine-tuning sur les tâches en aval.	Non
ESG-BERT [19]	Classification ESG (25 labels)	Pré-entraînement de BERT avec la tâche NSP sur des rapports annuels de développement durable et des articles d'actualité financière, puis fine-tuning sur des données labellisées.	Oui

Tableau 1 : Revue des modèles basés sur BERT et entrainés sur des données financières ou ESG [7]

III. Etude de cas

Sources et types de données

Les investisseurs font appel à l'IA et aux nouvelles technologies pour automatiser la collecte et l'analyse des données ESG pour les marchés développés et émergents. Les sources de données ESG non structurées, telles que les articles d'actualité, les rapports des banques multilatérales de développement (BMD) ainsi que les rapports annuels des entreprises, sont généralement sous-utilisées dans les analyses de la performance ESG des entreprises.

L'utilisation d'applications d'IA, renforcée par une nouvelle génération d'algorithmes d'apprentissage machine (ML) et de cloud computing, a récemment conduit à des innovations dans l'analyse de données textuelles non structurées à grande échelle. Par exemple, ChatGPT [26, 27] est un modèle entraîné sur de vastes volumes de données textuelles pour générer des réponses à des questions ou des dialogues de conversation.

Plusieurs sources de données sont disponibles selon le cas d'usage. Ci-dessous deux exemples de sources gratuites de données dynamiques qui peuvent être utilisées pour une première analyse ESG temporelle.

- GDELT [2] est une base de données qui surveille les actualités mondiales diffusées, sous forme papier et Web dans tous les pays, dans plus de 100 langues et identifie les personnes, les lieux, les organisations, les thèmes, les sources, les émotions, les citations, les images et les événements qui animent la société mondiale à chaque seconde de chaque jour. Cette base de données est actualisée chaque 15 minutes. Les actualités sont traduites en temps réel en anglais et analysées par des modèles internes qui produisent des scores de polarités et de sentiment. Néanmoins, la base de données GDELT ne donne pas un accès direct au contenu de l'article mais aux métadonnées des articles (les personnes, les lieux, les sentiments, les organisations …), ce qui limite l'utilisation de cette base dans notre approche d'analyse avec des modèles NLP.

- l'API New York Times [3] : New York Times propose une API appelée "Archive", cette API renvoie un tableau d'articles NYT pour un mois donné, remontant jusqu'à 1851. Ses champs de réponse incluent le résumé de la publication, le titre de la publication, le type de la publication, la section de publication, etc... L'API Archive est très utile pour collecter une base de métadonnées d'articles NYT avec un résumé de chaque publication. Dans cette étude de cas, on utilisera des données du New York Times pour tester les différents modèles NLP considérés.

Démarche

Dans cette section, on présente un cas d'usage qui démontre l'application du NLP dans l'analyse ESG en utilisant des données collectées à partir de l'API du New York Times [3]. Les données collectées incluent des articles de presse liés aux affaires et à la finance, que nous avons filtrés (comme décrit ci-dessous) pour fournir des informations pertinentes à notre étude. La Figure 2 présente la démarche suivie pour effectuer l'analyse.

Ce cas d'usage montre comment des techniques NLP telles que l'analyse de sentiment et la classification de texte, peuvent être utilisées pour obtenir des informations précieuses sur les tendances et les sentiments ESG. En utilisant la puissance du NLP, on vise à fournir une analyse complète des sentiments ESG pour différentes entreprises, qui peuvent ensuite être comparées aux scores ESG fournis par d'autres fournisseurs.

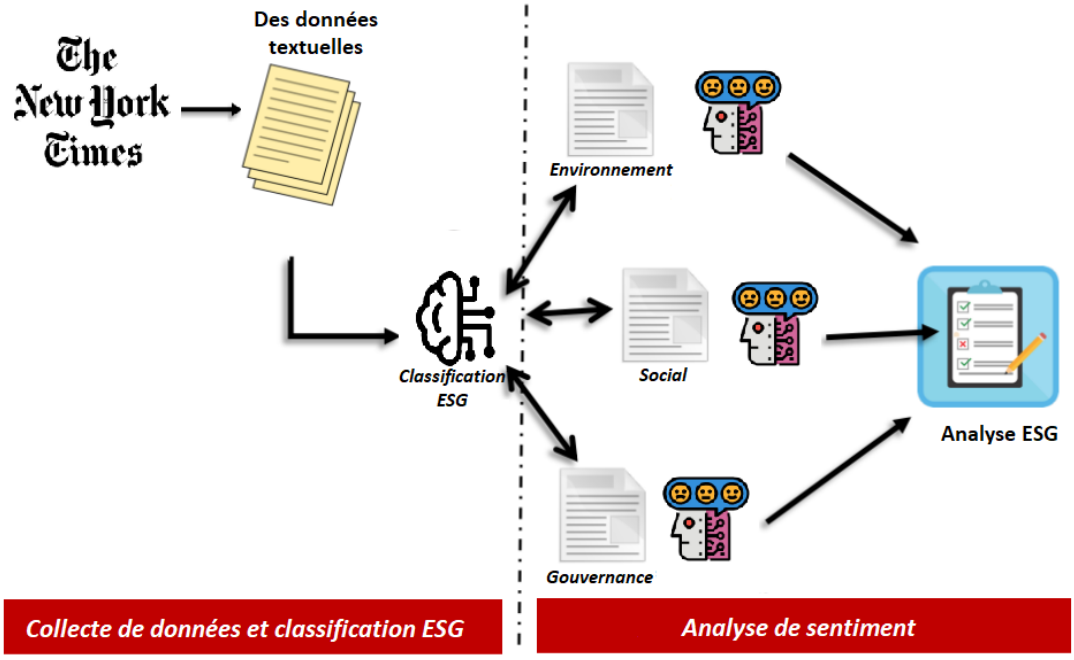
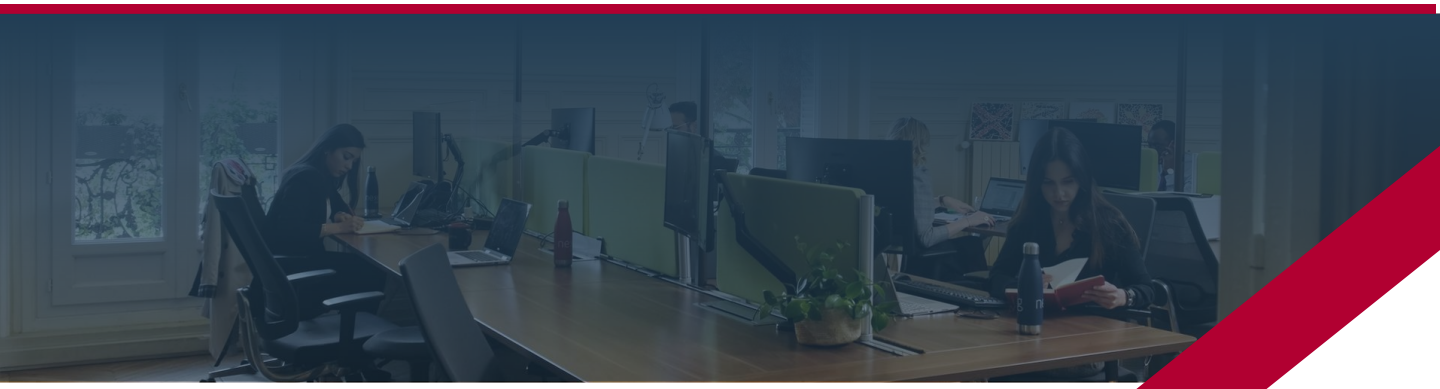
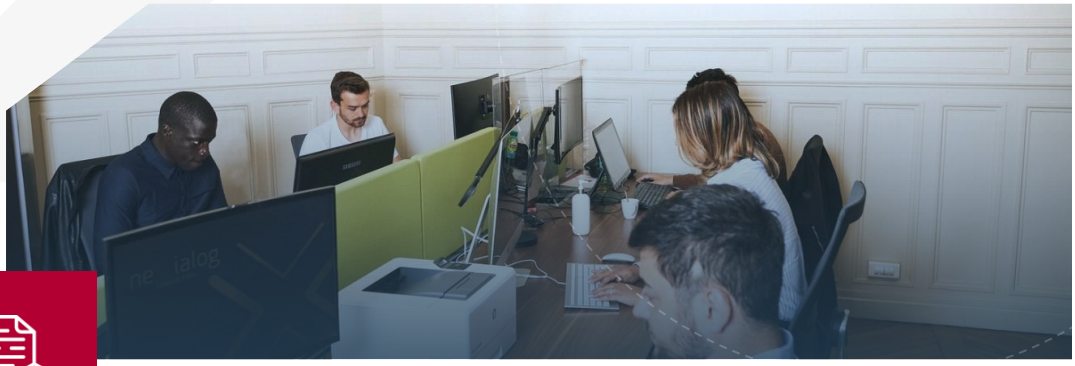


Figure 2 : Démarche suivie dans l'étude de cas





Préparation des données

L'API du New York Times permet de récupérer toutes les métadonnées de toutes les publications. Ces métadonnées contiennent des informations sur :

- Les organisations (entreprises) liées à la publication
- Le type de publication : Actualité, Éditorial, Revue…
- Les thèmes : Affaires, Économie, International, Politique…
- Le titre de la publication
- Le résumé de la publication

Le type, en plus des thèmes de la publication, permet de filtrer les données à récupérer. Ainsi on a collecté toutes les publications depuis 2018 jusqu'à 2023 avec les thèmes : Affaires, Économie, Technologie et Climat et les types : Actualité, Sunday Business et Éditorial. Le titre et le résumé sont ensuite concaténés et utilisés comme inputs aux modèles énumérés dans le Tableau 1. Les résultats des modèles sont traités et agrégés par entreprise pour ensuite élaborer une analyse individuelle.

Résultats

Dans cette partie, on présente les résultats de l'analyse qui résulte de deux principales étapes :

La première étape consiste à :

- Sélectionner les entreprises ayant la plus grande quantité de données afin d'effectuer une analyse par entreprise la plus représentative possible.
- Effectuer une classification ESG et une classification plus fine des publications en 9 catégories (Changement climatique, Capital naturel, Responsabilité du produit, Relations avec la communauté, Pollution et déchets, Capital humain, Éthique et valeurs d'entreprise, Gouvernance d'entreprise, Non ESG).

La deuxième étape consiste à :

- Effectuer une analyse sentimentale par classe E, S et G pour les entreprises sélectionnées. Le but étant de quantifier l'implication de ces entreprises dans les critères ESG à partir d'une analyse sentimentale de leurs actualités.

Étape 1 : Classification ESG

Après filtrage, nous avons sélectionné dans les données collectées 55.697 actualités sur différentes organisations. La Figure 3 montre les entreprises avec le plus grand nombre de mentions dans les données. Dans la deuxième partie de l'analyse on restreint l'analyse ESG à ces entreprises.

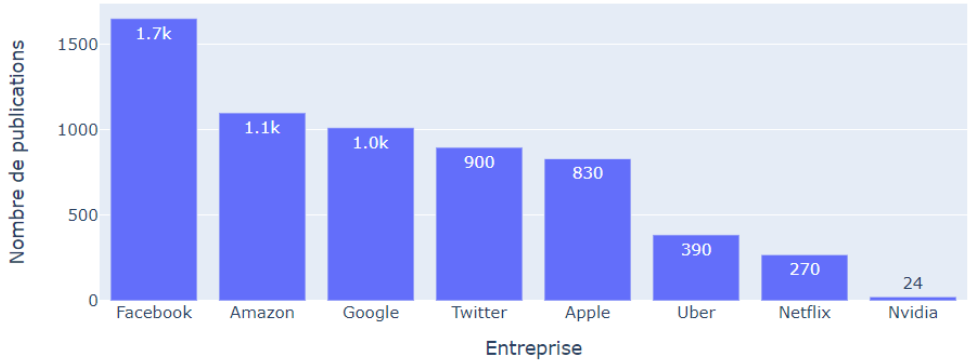


Figure 3 : Nombre de publications par entreprise

En utilisant le modèle FinBERT-ESG, on récupère les classes de chaque actualité, la Figure 4 montre le pourcentage de chaque catégorie.

Cette première classification permet non seulement de classifier les données en E, S et G, mais aussi de détecter et filtrer les données non-ESG. La Figure 4 montre la prédominance des actualités liées au critère Social avec un pourcentage élevé (63,8%). Les deux autres critères (E et G) sont moins bien représentés dans ces données.

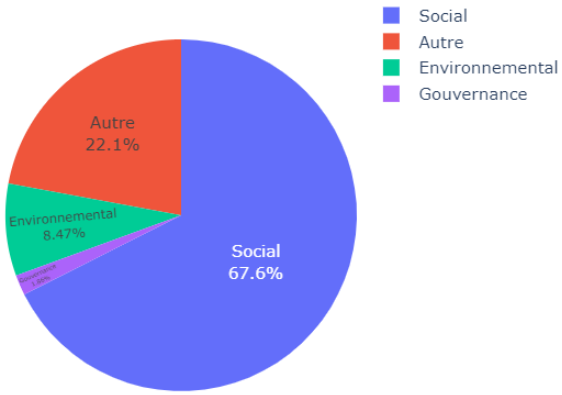


Figure 4 : Pourcentage des catégories E,S,G et Autre

Social : 37.637, Autre 12.309, Environnemental : 4.715, Gouvernance : 1.036

Pour affiner les catégories, on utilise le modèle *FinBERT-esg-9-catégories* qui classe chaque actualité en sous-catégories *(Changement climatique, Capital naturel, Responsabilité du produit, Relations avec la communauté, Pollution et déchets, Capital humain, Éthique et valeurs d'entreprise, Gouvernance d'entreprise, Non ESG)*, ceci permet d'effectuer une analyse plus fine pour chaque organisation sur les critères ESG. La Figure 5 montre les résultats de cette étape.

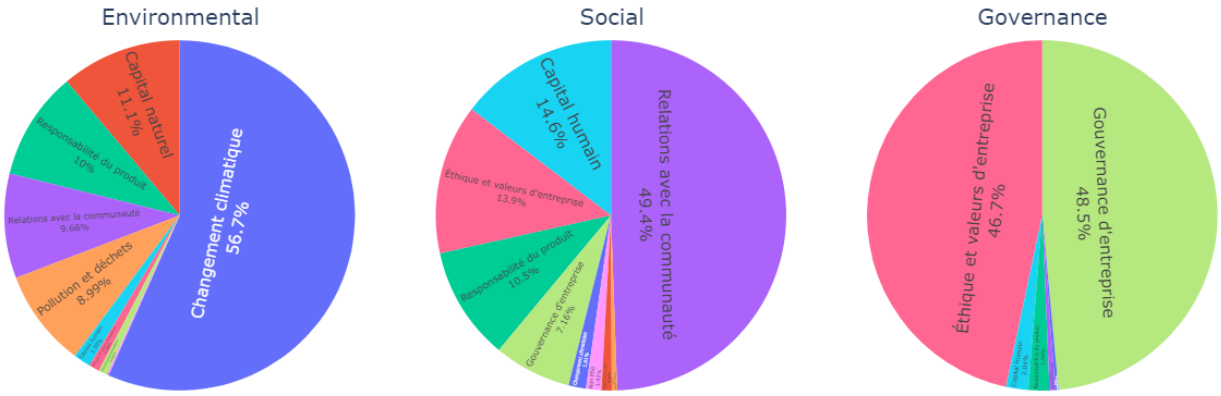
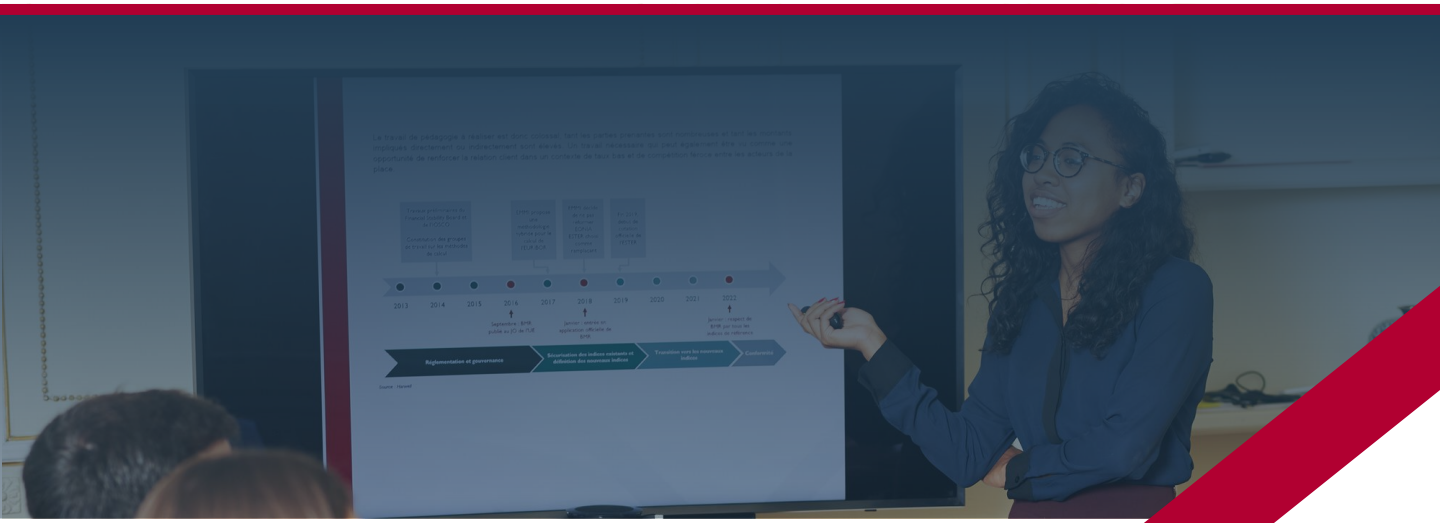


Figure 5 : Pourcentage de chaque sous-catégorie pour E,S et G

On remarque que les sous-catégories les plus présentes dans la classe "Environnement" sont "Changement Climatique" et "Capital Naturel", Pour la classe "Social", on trouve "Relations avec la communauté" et "Capital Humain", et finalement pour la classe "Gouvernance" les deux sous-catégories "Gouvernance d'Entreprise" et "Éthique et les valeurs d'entreprise" couvrent la majorité des actualités.

La partie suivante de l'analyse permet de mesurer, pour chaque entreprise retenue dans l'étude, les émotions et attitudes exprimées dans ces données classées E, S et G.



Étape 2 : Analyse de sentiment

La deuxième étape de l'analyse consiste à analyser les sentiments pour chaque organisation en fonction de la catégorie (E,S ou G) et analyser les scores sur une période de temps donnée. Le score de tonalité est calculé en normalisant la probabilité de positivité et négativité pour chaque publication :

$$Score\ de\ tonalité = \frac{Probabilité\ de\ positivité\ -\ Probabilité\ de\ négativité\ +\ 1}{2} \in [0,1]$$

Ce score est borné entre 0 et 1 et représente un spectre de sentiment allant du négatif en 0 au positif en 1 en passant par le neutre en 0.5.

La Figure ci-dessous montre pour chaque entreprise retenue la moyenne et l'écart type des scores de tonalité sur la période 2018 jusqu'à 2023. Ces scores sont calculés séparément pour chaque catégorie (E,S et G).

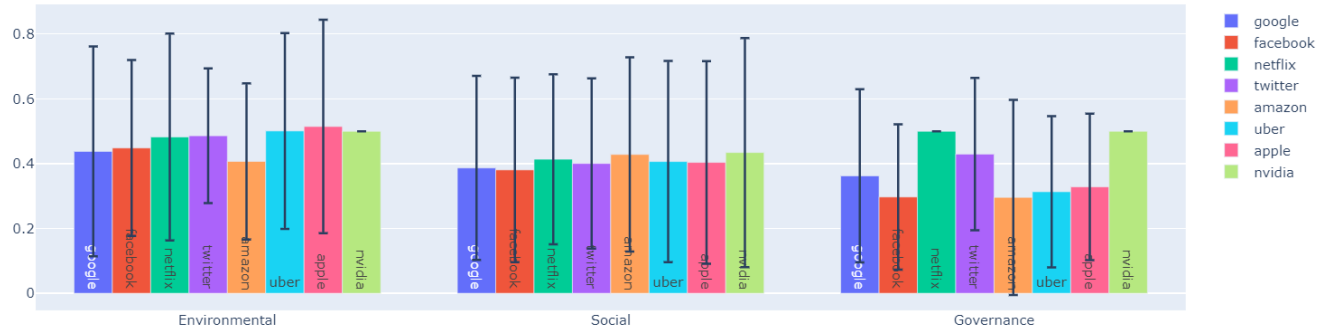


Figure 6 : moyenne et l'écart type des scores de tonalité ESG pour quelques entreprises.

La Figure 6 révèle une certaine incertitude sur l'estimation des scores sur la période choisie, ce qui traduit une instabilité des scores des entreprises sélectionnées sur cette période. À noter que l'écart type de Nvidia et Netflix est égal à 0 sur certains critères à cause du manque de données. Il est ainsi important de prendre en compte le nombre de publications disponibles pour chaque entreprise, ce nombre peut varier considérablement ce qui rend l'analyse des scores difficile et peut potentiellement biaiser les résultats. En effet, une entreprise qui a été le sujet de nombreuses publications dans une catégorie donnée, aura certainement un score plus représentatif dans cette catégorie que d'autres entreprises qui ont été moins couvertes par les médias. De même, une entreprise qui a peu été couverte par les médias peut se retrouver avec des scores moins représentatifs de la réalité, simplement en raison de l'absence d'actualités pertinentes (voir Tableau 2).

Le tableau suivant regroupe le nombre de publications utilisé pour construire les scores de la Figure 6 pour chaque organisation par catégorie.

Organisation	Environnemental	Social	Gouvernance	Total
Google	39	808	16	**863**
Facebook	39	1381	18	**1438**
Netflix	6	186	1	**193**
Twitter	23	683	13	**719**
Amazon	53	853	14	**920**
Uber	15	325	3	**343**
Apple	37	582	8	**627**
Nvidia	1	11	0	**12**
Total	**213**	**4829**	**73**	**5115**

Tableau 2 : Nombre de publication par organisation et par catégorie

Le tableau confirme le grand déséquilibre des classes E,S et G (voir également Figure 4) par entreprise dans ces données. Ainsi dans la suite, on focalise notre analyse sur la classe S qui regroupe le plus de données.

Pour réaliser une analyse temporelle sur le critère S, nous considérons les deux entreprises Facebook et Google qui sont les mieux représentées dans cet échantillon. La Figure 7 montre l'évolution des scores de tonalité entre 2018 et 2023.

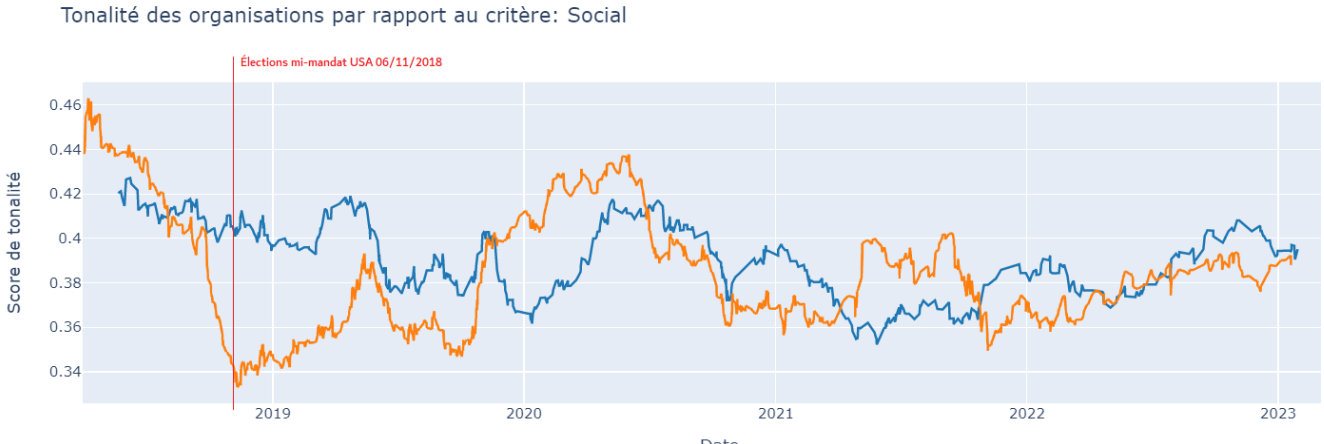
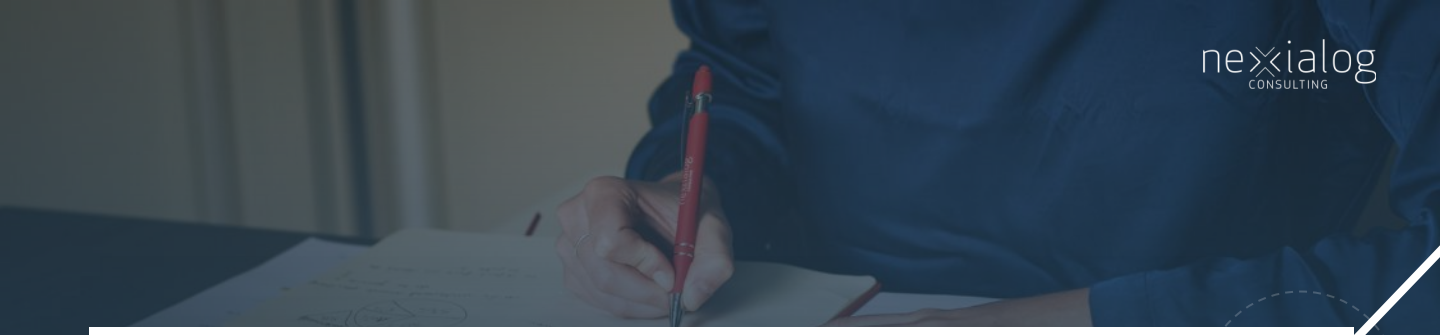


Figure 7 : Evolution du score de tonalité par rapport au critère 'Social' pour les deux entreprises : Facebook — et Google —. Une moyenne mobile sur une fenêtre de 200 jours a été appliquée pour lisser les courbes.



On remarque que les variations des scores , entre 2018 et 2023, pour les deux entreprises se situent dans une plage de valeurs similaires avec des moyennes très proches l'une de l'autre. On remarque que les scores ne dépassent pas 0.5 (neutre) sur la période pour les deux entreprises, ce qui est en phase avec les résultats de la Figure 6. Cependant, une baisse est constatée pour Facebook par rapport à Google pendant la période novembre 2018 à février 2019. Afin de mieux comprendre les raisons de cette chute, une analyse plus approfondie des données pour cette période est nécessaire.

Pendant cette période, la majorité des publications portent sur l'élection de mi-mandat aux États-Unis et d'autres sujets. Les thèmes les plus discutés sont regroupés dans la liste suivante :

- Le rôle de Facebook dans la propagation de la désinformation dans les campagnes politiques.
- Les violations de la vie privée de Facebook, y compris l'accès aux photos privées et le scandale Cambridge Analytica.
- L'implication de Facebook dans l'incitation à la violence au Myanmar.
- L'impact de Facebook sur la société et les opportunités d'emploi ainsi que son influence sur le marché boursier.

Limitations liées à l'utilisation des données

Cette note a pour but d'illustrer les potentiels des modèles NLP dans le contexte ESG sur un cas d'usage simple en utilisant la base de données du New York Times. Bien que cette source de données soit précieuse pour l'analyse ESG, il est important de reconnaître que son contenu n'est pas directement lié à l'ESG et on remarque un grand déséquilibre entre les classes E,S et G ce qui est complètement en phase avec le type de publication du New York Times. En plus, son contenu peut être influencé par des facteurs tels que la géographie et les tendances éditoriales. En outre, la base de données peut ne pas refléter la diversité des perspectives et des opinions sur les questions ESG, ce qui peut également introduire des biais dans l'analyse. Par conséquent, il est essentiel de compléter les données du New York Times avec d'autres sources pour obtenir une image plus complète et précise des performances ESG des entreprises.

Conclusion

En conclusion, cet article vise à illustrer l'utilisation des techniques du traitement du langage naturel (NLP) sur une base de données en libre accès pour construire un score ESG.

L'article présente plusieurs techniques NLP, telles que le topic modeling (modélisation thématique), la classification de texte et l'analyse de sentiment, qui peuvent être utilisées de manière combinée pour fournir des informations précieuses sur les tendances et les sentiments ESG.

Les techniques NLP sont illustrées sur un cas d'usage pratique qui démontre leur application dans l'analyse ESG à l'aide des données collectées à partir de l'API du New York Times.

Enfin, l'article souligne les risques associés à l'utilisation de ces modèles et l'importance des jeux de données utilisés lors de l'inférence.

L'utilisation du NLP dans l'analyse ESG offre un potentiel immense pour les investisseurs et les entreprises qui cherchent à comprendre les impacts environnementaux, sociaux et de gouvernance de leurs activités.

Bibliographie

[1] Conditions d'utilisation de l'API du New York Times, https://developer.nytimes.com/terms

[2] The GDELT project, https://www.gdeltproject.org/.

[3] New York Times « Archive API », https://developer.nytimes.com/docs/archive-product/1/overview.

[4] D. M. Blei et al. Latent dirichlet allocation. Journal of Machine Learning

Research, 3:993–1022, 2003, https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[5] Hofmann, Thomas. Probabilistic Latent Semantic Analysis. arXiv, 23 Jan. 2013, https://doi.org/10.48550/arXiv.1301.6705.

[6] Y. Whye Teh et al. Hierarchical Dirichlet Processes. EECS Berkeley, 15 Nov. 2005, https://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf.

[7] Devlin, Jacob, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv, 24 May 2019, https://doi.org/10.48550/arXiv.1810.04805.

[8] Liu, Yinhan, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv, 26 Juillet 2019, https://doi.org/10.48550/arXiv.1907.11692.

[9] DistilRoBERTa, Version distillée par HuggingFace en suivant la même demarche de DistilBERT, https://huggingface.co/distilroberta-base.

[10] Sanh, Victor, et al. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv, 29 Feb. 2020, https://doi.org/10.48550/arXiv.1910.01108.

[11] Lee, Jinhyuk, et al. "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining." Bioinformatics, vol. 36, no. 4, Feb. 2020, pp. 1234–40. arXiv.org, https://doi.org/10.1093/bioinformatics/btz682.

[12] Beltagy, Iz, et al. SciBERT: A Pretrained Language Model for Scientific Text. arXiv, 10 Sept. 2019. arXiv.org, https://doi.org/10.48550/arXiv.1903.10676.

[13] Gutiérrez-Fandiño, Asier, et al. FinEAS: Financial Embedding Analysis of Sentiment. arXiv, 19 Nov. 2021. arXiv.org, https://doi.org/10.48550/arXiv.2111.00526.

[14] DistilRoberta Finetuned on financial sentimental analysis. 2022. HuggingFace https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis.

[15] A. Hazourli. FinancialBERT - a pretrained language model for financial text mining. 2022. https://www.researchgate.net/publication/358284785.

[16] Loukas, Lefteris, et al. FiNER: Financial Numeric Entity Recognition for XBRL Tagging. arXiv, 19 Apr. 2022. arXiv.org, https://doi.org/10.48550/arXiv.2203.06482.

[17] Huang, Allen, et al. FinBERT - A Large Language Model for Extracting Information from Financial Text. 28 Juillet 2020. Social Science Research Network, https://doi.org/10.2139/ssrn.3910214.

Bibliographie

[18] Mehra, Srishti, et al. "ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices." Embedded Systems and Applications, 2022, pp. 183–90. arXiv.org, https://doi.org/10.5121/csit.2022.120616.

[19] Mukut Mukherjee, et al. "ESG-BERT: Domain Specific BERT Model for Text Mining in Sustainable Investing" 2020. https://towardsdatascience.com/nlp-meets-sustainable-investing-d0542b3c264b.

[20] FactSet, spécialisée dans la fourniture de données et d'outils d'analyse financière pour les professionnels de la finance. https://www.factset.com.

[21] Datamaran, plateforme d'analyse de risques et d'opportunités ESG basée sur l'intelligence artificielle pour les entreprises. https://www.datamaran.com.

[22] Entis, plateforme qui utilise l'intelligence artificielle pour aider ses clients en générant des informations exploitables pour l'investissement. https://entis.ai.

[23] MSCI, spécialisée dans la fourniture d'outils d'analyse et d'indices boursiers (dont l'ESG) utilisés par les investisseurs pour évaluer les performances de leurs investissements. https://www.msci.com/.

[24] Sesamm, une société d'intelligence artificielle au service des sociétés d'investissement. https://www.sesamm.com/.

[25] RavenPack est un fournisseur de données volumineuses pour les services financiers, https://www.ravenpack.com/

[26] Le site officiel de ChatGPT (OpenAI) https://openai.com/blog/chatgpt

[27] Brown, Tom B., et al. Language Models Are Few-Shot Learners. arXiv, 22 Juillet 2020. arXiv.org, https://doi.org/10.48550/arXiv.2005.14165.

[28] Stephanie Mooij, The ESG Rating and Ranking Industry. The Issue of [ESG] Reporting Fatigue, Juin 2017, https://www.researchgate.net/publication/317687172_The_ESG_Rating_and_Ranking_Industry_The_Issue_of_ESG_Reporting_Fatigue

[29] Philippe Aubain, EY France, Reporting ESG des entreprises françaises : sont-elles prêtes pour la CSRD?, Nov 2022 https://www.ey.com/fr_fr/climate-change-sustainability-services/reporting-esg-des-entreprises-francaises-pretes-pour-la-csrd

[30] Reimers, Nils, and Iryna Gurevych. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. arXiv, 27 Aug. 2019. arXiv.org, https://doi.org/10.48550/arXiv.1908.10084.

Annexe

Cette annexe regroupe les détails de pré-entraînement et fine-tuning des modèles présentés dans le Tableau 1. Les démarches diffèrent pour chaque modèle et chaque tâche. Nous présentons un résumé pour chaque modèle et chaque tâche :

Avant de lister les modèles. Nous présentons une base de données qui est utilisée par la plupart des modèles dans la phase de fine-tuning :

- **Financial PhraseBank** est une base de données de 4840 nouvelles financières classées par sentiment (négatif, neutre, positif), utilisée souvent pour effectuer le fine-tuning des modèles.

Les démarches d'entraînement des modèles sont expliquées ci-dessous :

1. **FinEAS** :

Ce modèle est une version fine-tuned de sentence-BERT (sans pré-entraînement) sur des données annotées de RavenPack.

Sentence-BERT: est un modèle BERT qui est fine-tuned en supervision avec la technique de siamese BERT network [30], Cette approche aboutit à des représentations de phrases plus significatives par rapport au modèle BERT.

2. **DistilRoberta Finetuned** :

DistilRoberta Finetuned est une version fine-tuned avec la base de données Financial PhraseBank du modèle DistilRoberta sur la tâche de l'analyse de sentiment.

3. **FinancialBERT** :

FinancialBERT est un modèle BERT pré-entraîné avec les deux techniques MLM et NSP sur un large corpus de textes financiers de 3.39 milliards de tokens (TRC2-financial: 0.29 milliards de tokens, Bloomberg News 0.2 milliards de tokens, Corporate Reports: 2.2 milliards de tokens, Earnings Call Transcripts: 0.7 milliards de tokens)

- FinancialBERT-Sentiment-Analysis : est un modèle FinancialBERT fine-tuned sur la base de données Financial PhraseBank. Ce modèle permet ainsi d'accomplir une analyse sentimentale des phrases.

4. **SEC-BERT** :

SEC-BERT est un modèle BERT pré-entraîné sur une base de données de 260 773 rapports 10-K de 1993 à 2019, accessibles auprès de "U.S. Securities and Exchange Commission (SEC)". Ce modèle est ensuite fine-tuned sur l'ensemble de données de Financial Phrasebank et un jeu de données Kaggle qui comprend des données de sentiment Covid-19.

Annexes & références

5. **FinBERT** :

FinBERT est un modèle BERT pré-entraîné avec les deux techniques MLM et NSP sur un corpus de communication financière d'une taille totale de 4,9 milliards de tokens (Rapports d'entreprise 10-K & 10-Q : 2,5 milliards de tokens, Transcriptions d'appels sur les revenus : 1,3 milliard de tokens et Rapports d'analystes : 1,1 milliard de tokens).

- *FinBERT-ESG* : est un modèle FinBERT fine-tuned sur 2000 phrases annotées manuellement à partir des rapports ESG et des rapports annuels des entreprises. Le modèle permet de classifier un texte donné en input en trois catégories : Environnement, Social, Gouvernance.
- *FinBERT-esg-9-catégories* : est un modèle FinBERT fine-tuned sur environ 14000 phrases annotées manuellement à partir des rapports ESG et des rapports annuels des entreprises. Ce modèle classe un texte en neuf sujets ESG détaillés: ENV (changement climatique, capital naturel, pollution et déchets), SOC(capital humain, responsabilité du fait des produits, relations communautaires), GOV(gouvernance d'entreprise, éthique des affaires et valeurs), et non ESG. Ce modèle complète finbert-esg qui classe un texte en quatre thèmes ESG (E, S, G ou Aucun).
- *FinBERT-tone* : est un modèle FinBERT fin-tuned sur 10000 phrases annotées manuellement (positive, négative, neutre) à partir des rapports d'analystes. Ce modèle permet ainsi d'accomplir une analyse sentimentale des phrases. L'article présente une comparaison avec BERT et d'autres modèles de Machine Learning sur la base Financial Phrasebank.

6. **ESGBERT** :

ESGBERT est un modèle BERT pré-entraîné par des des données collectées à partir du "Knowledge Hub" du projet "Accounting for Sustainability". Ce modèle est ensuite fine-tuned sur des rapports 10-Q des entreprise S&P 500 annotés par scores obtenus à partir de la platforme de recherche de Wharton (WRDS).

7. **ESG-BERT** :

ESG-BERT a été entraîné sur des données textuelles non structurées avec des précisions de 100 % et 98 % pour les tâches de NSP et MLM. Le fine-tuning d'ESG-BERT pour la classification du texte a donné un score F-1 de 0.9. À titre de comparaison, le modèle général BERT (BERT-base) a obtenu un score de 0.79 après fine-tuning.

Nexialog Consulting

STRATÉGIE | ACTUARIAT | GESTION DES RISQUES | DATA

Nexialog Consulting est un cabinet de conseil spécialisé en Stratégie, Actuariat, Gestion des risques et Data qui dessert aujourd'hui les plus grands acteurs de la banque et de l'assurance. Nous aidons nos clients à améliorer de manière significative et durable leurs performances et à atteindre leurs objectifs les plus importants.

Les besoins de nos clients et les réglementations européennes et mondiales étant en perpétuelle évolution, nous recherchons continuellement de nouvelles et meilleures façons de les servir. Pour ce faire, nous recrutons nos consultants dans les meilleures écoles d'ingénieur et de commerce et nous investissons des ressources de notre entreprise chaque année dans la recherche, l'apprentissage et le renforcement des compétences.

Quel que soit le défi à relever, nous nous attachons à fournir des résultats pratiques et durables et à donner à nos clients les moyens de se développer.

CONTACTS

Ali BEHBAHANI
Associé, Fondateur
+ 33 (0) 1 44 73 86 78
abehbahani@nexialog.com

Christelle BONDOUX
Associée, Direction Commerciale & Recrutement
+ 33 (0) 1 44 73 75 67
cbondoux@nexialog.com

Vivien BRUNEL
Associé, Data & Innovation
vbrunel@nexialog.com

Paul-Antoine DELETOILLE
Sales Leader
+33 (0)1 44 73 75 70
+33 (0)7 64 57 86 69
padeletoille@nexialog.com

Areski COUSIN
Directeur scientifique
acousin@nexialog.com

www.nexialog.com

Retrouvez toutes nos publications sur Nexialog R&D