Texte comme donnée en Finance: Entre signal et bruit

Journée Audaces 2025
Intelligence Artificielle
5 juin 2025

Selim MANKAÏ
Département Finance, IAE, UCA

CONTEXTE

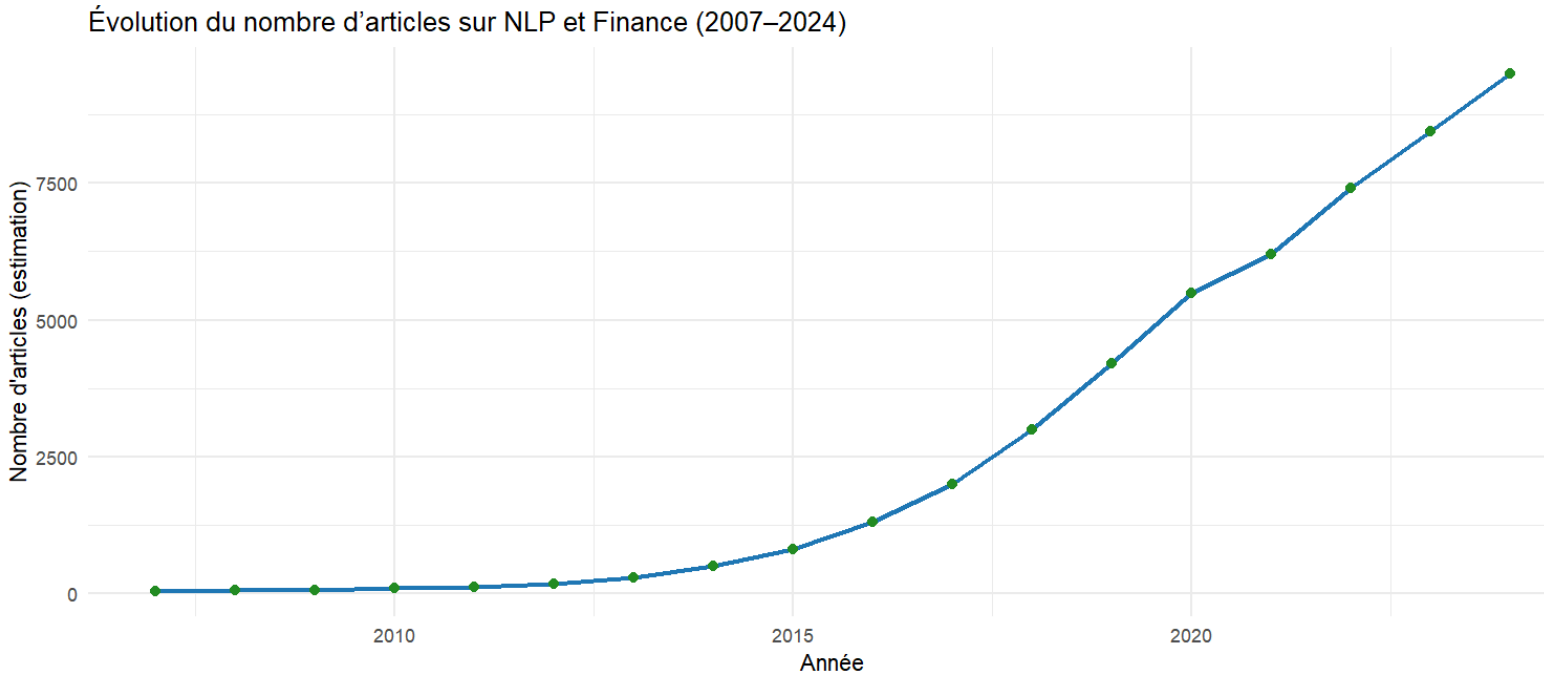
- Plus de 5,4 milliards d'individus connectés à Internet, des interactions sociales se déroulent de plus en plus par écrit (Statista, 2024).
- Chaque jour, une personne moyenne échange environ 80 courriels et 50 messages textuels (GilPress, 2024).
- Une personne parcourt environ 90 mètres de contenu en ligne, soit l'équivalent de trois fois le « New York Times » (Robertson et al, 2024).

CONTEXTE

- Plus de 5,4 milliards d'individus connectés à Internet, des interactions sociales se déroulent de plus en plus par écrit (Statista, 2024).
- Chaque jour, une personne moyenne échange environ 80 courriels et 50 messages textuels (GilPress, 2024).
- Une personne parcourt environ 90 mètres de contenu en ligne, soit l'équivalent de trois fois le « New York Times » (Robertson et al, 2024).
- 85 à 90 % des données d'entreprise sont aujourd'hui non structurées (e.g., reporting, e-mails,, images, vidéos et publications sur les réseaux sociaux, etc.) (KNIME, 2021)
- Le volume de ces données double tous les 18 à 24 mois, représentant un défi majeur pour leur gestion et leur exploitation.
- Généralisation du format XBRL (eXtensible Business Reporting Language) les rapports financiers aux USA (SEC, 2009)

EVOLUTION EXPONNTIONNELLE

- Multiplications des méthodes
- Exposition des recherches mobilisant le NLP en Finance



ENJEUX

- Données précieuses sur les comportements, les opinions et les perceptions des entreprises.
- Nature non structurée des données et complexité de l'analyse.
- Pertinence des méthodes de text mining et le traitement automatique du langage naturel (NLP) pour extraire des informations pertinents.
- Nouvelles contributions empiriques majeures en SHS

QUELLES CONNAISSANCES PRODUIT-ON ?

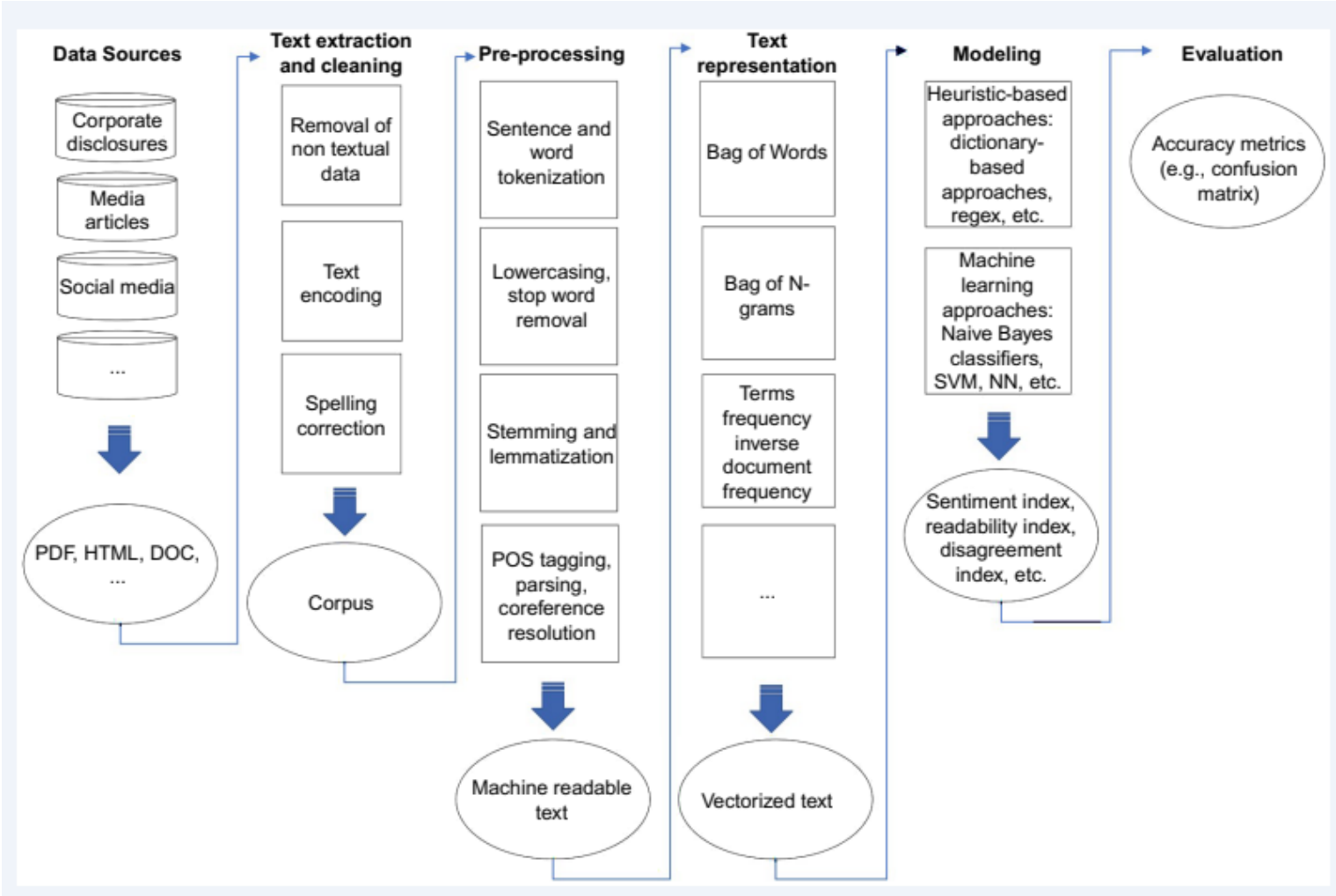
Ces tendances soulèvent des questions épistémologiques:

- Que mesure-t-on réellement?
- Comment distinguer le signal pertinent du bruit discursif ?

PLAN DE LA PRÉSENTATION

- Bref descriptif des méthodes NLP en finance
- Présentation de trois cas (articles de recherche)
- Synthèse autour des notions de signal/bruit.

NLP PIPELINE



Corazza, Marco, et al.. Artificial intelligence and beyond for finance. World scientific, 2024.

DONNÉES FINANCIÈRES (STRUCTURÉES)

- Ratios comptables (ex : rentabilité, liquidité, endettement…)
- États financiers (bilan, compte de résultat, cash flow)
- Données boursières : prix, volumes, volatilité, rendement
- Notations de crédit, probabilités de défaut

Sources : Compustat, CRSP, Bloomberg, Refinitiv, Orbis, Moody's

Usage :Prédiction de défaut, valorisation d'actifs, scoring, stress testing…

DONNÉES TEXTUELLES

- Documents réglementés:

Rapport annuel 10-K (USA); Form 10-Q, 8-K; Prospectus, déclarations aux autorités (SEC, AMF…)

- Données textuelles non réglementées (médias & opinions):

Articles de presse; Blogs financiers, newsletters d'analystes; Forums spécialisés (ex : Seeking Alpha)

- Données sociales & comportementales:

Tweets, Reddit, enquêtes

- Données qualitatives internes:

Transcriptions de réunions (earnings calls, AG); Comptes-rendus de conseil d'administration; Communication volontaire ESG/RSE

REPRÉSENTATION DU TEXTE

- Bag of Words (BoW)

Transforme un texte en un vecteur de fréquences brutes des mots.

- TF-IDF

Mesure de la rareté du mot dans l'ensemble du corpus

- Word2Vec

Représente chaque mot comme un vecteur dense appris à partir de cooccurrences.

- Contextual Embeddings (ELMo, BERT)

Embeddings contextuels : un mot a un vecteur différent selon la phrase

- Sentence Embeddings (SBERT)

Représente des phrases/documents sous forme de vecteurs sémantiques

- Language Models Génératifs (GPT, T5, BART)

Comprendre, compléter, traduire ou résumer un texte.

QUELQUES FINALITES DES MODLES

Les méthodes sont différenciées:

Logique méthodologique (déductive vs inductive)
Objectif analytique (mesurer, explorer, prédire, extraire...)
Niveau technique

- Dictionnaires → interprétation simple, comptage ciblé
- Topic modeling → émergence de thèmes
- Transformers → réponse contextuelle, génération de textes

APPROCHES HYBRIDES ET ADAPTATIVES

Approche en deux temps : représentation vectorielle (Word2Vec, BERT) suivie d'un modèle prédictif.

Chaîne de traitement modulaire : extraction des caractéristiques → modélisation → interprétation.

Des croisements de logiques : inductif (exploration) déductif (test d'hypothèse).

Emergence de modèles avancés, RAG (Retrieval-Augmented Generation) ou les mixtures d'experts (MoE).

EXEMPLE 1 : CHAMP D'ACTIVITÉ LATENT

The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

Original Article

Scope, Scale, and Concentration: The 21st-Century Firm

GERARD HOBERG, GORDON M. PHILLIPS

First published: 04 November 2024 | https://doi.org/10.1111/jofi.13400 | Citations: 7

Read the full text > PDF TOOLS SHARE

ABSTRACT

We provide evidence using firm 10-Ks that over the past 30 years, U.S. firms have expanded their scope of operations. Increases in scope were achieved largely without increasing traditional operating segments. Scope expansion significantly increases valuation and is realized primarily through acquisitions and investment in R&D, but not through capital expenditures. Traditional concentration ratios do not capture this expansion of scope. Our findings point to a new type of firm that increases scope through related expansion, which is highly valued by the market.

EXEMPLE 1 : MOTIVATION

Sur les 30 ans, les entreprises américaines ont fortement élargi leur champ d'activité (scope) sans que cela apparaisse dans les mesures traditionnelles (segments Compustat).

Les mesures classiques (segments principaux déclarés) sous-estiment l'étendue réelle des entreprises.

Cela fausse la lecture sur la concentration des marchés (SFAS 131, 1997), la croissance des firmes, et peut mener à des erreurs en matière de régulation de la concurrence.

Entreprise	Segments Compustat
Coca-Cola	1 segment ("Boissons")
General Electric	3 segments ("Santé", "Énergie", "Aviation")

EXEMPLE 1 : SEGMENTS COMPUSTAT

Type de segment	Exemples d'intitulés déclarés dans Compustat
Technologie	Cloud Services, Software, Hardware, Internet of Things
Santé / Pharma	Pharmaceuticals, Medical Devices, Biotechnology
Énergie	Oil & Gas Production, Renewable Energy, Downstream Operations
Industrie	Aerospace, Construction Equipment, Industrial Automation
Finance / Assurance	Insurance, Investment Banking, Wealth Management
Consommation	Apparel, Food & Beverage, Retail, E-commerce
Transport / Logistique	Freight, Passenger Transport, Logistics Services
Télécommunications	Mobile, Fixed Line, Media Services

EXEMPLE 1 : OBJECTIF

1. Créer de nouvelles mesures de "scope" (étendue de marché) basées sur le texte des rapports annuels (10-K) via des techniques (NLP).
2. Montrer que ces mesures révèlent une expansion invisible du champ d'action.
3. Analyser les conséquences économiques de cette expansion.

CORPUS DE DONNÉES

Rapports annuels 10-K, section Item 1 – Business

Unité d'analyse : firme-année

Nombre total d'observations : 101 535 firmes-années

Période couverte :1989 à 2017

Sources:

EDGAR (SEC) : rapports électroniques (à partir de 1996)

Archives papier (Harvard, Dartmouth) : rapports scannés (1989–1995)

NLP PIPILINE

1. Prétraitement et nettoyage

   Suppression du contenu générique; Stop words (sans valeur sémantique)

2. Vectorisation avec Doc2Vec:

   Convertir chaque texte en un vecteur dense de 300 dimensions.

3. Clustering des segments industriels (D2V-industries)

   a. Sélection des firmes "mono-segment« (base d'apprentissage)

   b. Clustering (k-means)

   c. Élimination des clusters génériques

4. Attribution des segments aux firmes (Scorage)

   a. Calcul de la similarité (mesure cosinus)

   b. Seuil de pertinence

   c. Définition du scope

VALIDATION

Économiquement significative (Prédiction de la performance financière)

Plus informative que les mesures existantes (segments Compustat, NAICS)

Capable de capturer la diversité réelle des activités d'une entreprise

EXEMPLE 2: EXPOSIITON AU CHANGEMENT CLIMATIQUE

The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXVIII, NO. 3 • JUNE 2023

Firm-Level Climate Change Exposure

ZACHARIAS SAUTNER, LAURENCE VAN LENT, GRIGORY VILKOV, and RUISHEN ZHANG*

ABSTRACT

We develop a method that identifies the attention paid by earnings call participants to firms' climate change exposures. The method adapts a machine learning keyword discovery algorithm and captures exposures related to opportunity, physical, and regulatory shocks associated with climate change. The measures are available for more than 10,000 firms from 34 countries between 2002 and 2020. We show that the measures are useful in predicting important real outcomes related to the net-zero transition, in particular, job creation in disruptive green technologies and green patenting, and that they contain information that is priced in options and equity markets.

OBJECTIFS

- Extraire une mesure textuelle de l'attention au changement climatique (*CCExposure*) de l'entreprise (i) à la date (t) selon trois dimensions :

- Opportunités (ex. : transition verte, énergie renouvelable)
- Risques réglementaires (ex. : taxe carbone, lois environnementales)
- Risques physiques (ex. : catastrophes naturelles, inondations)

- Lier cette mesure à des comportements réels des entreprises (innovation, emploi, performance boursière)
- Tester la valeur prédictive et l'interprétation de la mesure comme signal économique

CORPUS DE DONNÉES

Transcriptions d'appels de résultats ("earnings calls")

- 102895 transcripts
- 11197 entreprises cotées (55 pays)

Période couverte :2002 à 2020

Earnings calls: Téléconférences où les dirigeants des entreprises cotées présentent les résultats financiers récents, les perspectives futures et répondent aux questions des analystes financiers.

NLP PIPILINE

1. Définition d'un vocabulaire initial ("seed bigrams")
   (eg. : "climate change", "carbon tax") .

2. Découverte automatique de mots-clés (King, Lam & Roberts (2017).
   Recherche d'autres bigrams corrélés aux bigrams initiaux.

3. Classification thématique des bigrams :
   Général– CCExposure
   Opportunités – CCExposureOppRisques
   Réglementaires– CCExposureRegRisques
   Physiques– CCExposurePhy

4. Construction des mesures d'exposition

$$CCExposure_{i,t} = \frac{1}{B_{i,t}} \sum_{b \in B_{i,t}} \mathbf{1}[b \in C]$$

VALIDATIONS MULTIPLES (ROBUSTESSE)

1. Validation de contenu

   Audit humain

2. Validation convergente (ou externe)

   Brevets verts (green patents)

   Offres d'emploi vertes (green jobs)

   Émissions de CO_2

3. Validation discriminante/structurelle

   Décomposition de la variance

   Analyse AR(1) $\rightarrow$ persistance temporelle ($\rho = 0{,}61$)

4. Validation prédictive/économique

   Effet sur les rendements boursiers

EXEMPLE 3: PREDICATION DEFAILLANCE

Annals of Operations Research
https://doi.org/10.1007/s10479-025-06574-z

ORIGINAL RESEARCH

Check for updates

Business Failure Prediction From Textual and Tabular Data With Sentence-Level Interpretations

Henri Arno[1] · Klaas Mulier[2] · Joke Baeck[2] · Thomas Demeester[1]

Received: 1 March 2024 / Accepted: 7 March 2025

Abstract
Business failure prediction models are crucial in high-stakes domains like banking, insurance, and investing. In this paper, we propose an interpretable model that combines numerical and sentence-level textual features through a well-known attention mechanism. Our model demonstrates competitive performance across various metrics, and the attention weights help identify sentences intuitively linked to business failure, offering a form of interpretability. Furthermore, our findings highlight the strength of traditional financial ratios for business failure prediction while textual data—particularly when represented as keywords—is mainly useful to correctly classify corporate disclosures where the possibility of failure is explicitly mentioned.

CORPUS DE DONNÉES

- Rapports 10-K de sociétés cotées aux États-Unis
- Années : 2002 à 2019

Deux types de données extraites :

a. Ratios financiers (30 ratios standards) :

- Rentabilité, endettement, liquidité, activité

b. Texte (section MD&A)Moyenne :

6 800 mots par document

Nettoyé, découpé en phrases (~300 à 500)

REPRÉSENTATION DES TEXTES

TF-IDF (keywords)

Mots-clés (unigrammes et bigrammes) extraits après nettoyage

Embeddings de documents

Vecteurs denses produits par des modèles pré-entraînés (text-embedding-ada-002)

Embeddings de phrases (approche principale du modèle proposé)

Utilisation de sentence embeddings (all-MiniLM-L6-v2, HuggingFace).

MODÈLE AVEC ATTENTION

Objectif : identifier automatiquement quelles phrases sont les plus informatives pour prédire une faillite.

Certaines phrases expriment des risques réels (signal)

D'autres relèvent du langage standardisé, légal, marketing (bruit).

Combiner des ratios financiers (features numériques) et des représentations de phrases.

Développement d'un mécanisme d'attention qui pondère chaque phrase selon sa pertinence

VALIDATION

1. Séparation temporelle des données

   Entraînement : entreprises entre 2002 et 2011

   Validation (tuning) : entreprises de 2012 à 2015

   Test final : entreprises après 2015

2. Gestion du déséquilibre des classes

   Comme la faillite est un événement rare, sur-échantillonnage des cas de faillite dans le jeu d'entraînement

3. Évaluation finale sur le jeu de test (post-2015)

   AUC (Area Under ROC Curve)

SYNTHESE

- Une bonne problématique permet de justifier les choix de modèles, des filtres, des pipelines.

- Approche multiméthodes et hybride

- Validations Multiples (Robustesse)