UQAR
Université du Québec
à Rimouski

Application de l'apprentissage automatique à la veille technologique

Mémoire présenté
dans le cadre du programme de maîtrise en informatique
en vue de l'obtention du grade de maître ès sciences (M.Sc.)

PAR
© FRANCK NKOLONGO TSHIBANDA

Janvier 2025

Composition du jury :

Steven Pigeon, président du jury, Université du Québec à Rimouski

Mehdi Adda, directeur de recherche, Université du Québec à Rimouski

Said Echchakoui, co-directeur de recherche, Université du Québec à Rimouski

Bruno Bouchard, examinateur externe, Université du Québec à Chicoutimi

Dépôt initial le 13 septembre 2024 Dépôt final le 07 janvier 2025

I dedicate this to my wife, mother and father. Thanks for always being there for me.

AVANT-PROPOS

La rédaction de ce mémoire marque la fin de plusieurs mois de travail acharné et de dévouement dans le domaine de la surveillance technologique et de la recherche. Ma passion pour les nouvelles technologies, en particulier les méthodes avancées de traitement du langage naturel et l'apprentissage automatique, m'a incité à étudier les processus d'innovation dans le secteur du plastique. L'objectif de ce mémoire est de proposer une vision nouvelle et de contribuer à l'évolution des méthodes de surveillance technologique.

Je tiens tout d'abord à exprimer ma sincère reconnaissance envers mon directeur de recherche, le Dr Mehdi Adda, pour son grand soutien, ses conseils avisés tout au long de cette aventure académique. Grâce à ses commentaires constructifs et à son expertise, j'ai pu structurer et guider mes recherches.

Je souhaite également exprimer ma sincère gratitude envers mon codirecteur de recherche, Said Echchakoui, pour son précieux soutien et ses recommandations éclairées. Sa compétence et son engagement ont joué un rôle essentiel dans la création et la réalisation de ce mémoire.

Je tiens à exprimer ma gratitude envers ma famille, pour leur amour et leur soutien sans faille. Grâce à leur patience et à leur compréhension, je puis me consacrer entièrement à cette étude.

J'espère que cette étude pourra constituer le point de départ des études futures et apporter une contribution importante à l'évolution de la surveillance technologique et de l'innovation.

RÉSUMÉ

Dans ce mémoire, nous examinons une nouvelle approche méthodologique pour l'application de l'apprentissage automatique à la veille technologique en utilisant les informations provenant de différentes sources, comme les bases de données de brevets, les réseaux sociaux et les bases de données spécialisées. L'objectif de cette recherche est de saisir les dynamiques technologiques et de repérer les opportunités d'innovation dans divers secteurs industriels comme l'industrie du plastique, en utilisant des techniques avancées de traitement automatique du langage naturel (TAL) et des modèles pré-entraînés tels que RoBERTa. Spécifiquement, l'étude se concentre sur la collection et la représentation thématique des données textuelles, en incluant des mesures de similitude afin de repérer les tendances et les avancées. La méthode suggère une approche automatisée pour repérer les technologies brevetées et les opportunités technologiques en tirant parti des différences entre les données structurées (brevets) et non structurées (réseaux sociaux, etc.). Les résultats montrent que cette méthode renforce la surveillance technologique et assiste les entreprises dans l'anticipation des évolutions technologiques, en leur offrant des renseignements essentiels sur les tendances et les avancées. Les conclusions du mémoire portent sur les conséquences pratiques, les contraintes des méthodes actuelles et les perspectives de recherche à venir.

Mots clés : Veille technologique, Apprentissage automatique, Innovation, Opportunités, Industrie du plastique, LDA, BERT.

ABSTRACT

This thesis examines a new methodological approach for applying machine learning to technological monitoring using information from various sources, such as patent databases, social networks, and specialized databases. This research aims to grasp technological dynamics and identify innovation opportunities in different industrial sectors such as the plastics industry, using advanced natural language processing (NLP) techniques and pre-trained models like RoBERTa. Specifically, the study focuses on the collection and thematic representation of textual data, including similarity measures to identify trends and advancements. The method suggests an automated approach to identify patented technologies and technological opportunities by leveraging the differences between structured data (patents) and unstructured data (social networks, etc.). The results show that this method enhances technological monitoring and assists companies in anticipating technological developments, providing them with essential information on trends and advancements. The conclusions of the thesis focus on the practical implications, the limitations of current methods, and the prospects for future research.

Keywords: Technological Surveillance, Machine Learning, Innovation, Opportunities, Plastics Industry, LDA, BERT.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

NLP	Natural Langage Preprocessing
BERT	Bidirectional Encoder Representations from Transformers
LDA	Latent Dirichlet Allocation
UQAR	Université du Québec à Rimouski
SVM	Support Vector Machine
WoS	Web of Science
TOA	Technology Opportunity Analysis
SOA	Subject Object Action
TF	Term Frequency
IDF	Inverse Document Frequency
TAL	Traitement automatique de langage
RDI	Recherche développement et innovation
LLM	Large Language Model
TIC	Technologies de l'information et de la communication
API	Application Programming Interface

CHAPITRE 1 : INTRODUCTION GÉNÉRALE

1.1. CONTEXTE DE RECHERCHE ET PROBLÉMATIQUE

1.1.1. Contexte de recherche

La veille technologique est essentielle pour anticiper les avancées technologiques et élaborer des scénarios, des plans d'action et des stratégies dans un domaine précis. Les pays et les organisations suivent attentivement ces avancées pour formuler des stratégies et des politiques, identifiant ainsi les trajectoires technologiques futures (Kostoff & Schaller, 2001). Selon Daniel San *et al*. (2021), le processus consiste à rechercher, extraire, sélectionner, analyser, ajouter de la valeur et distribuer des informations, ce qui est considéré comme répétitif et improductif jusqu'à 65% du temps. Les technologies de l'information et de la communication automatisent certains aspects, réduisant ainsi les coûts de main-d'œuvre humaine (Perez *et al.*, 2018). Selon le Dr Alan Porter (Porter & Detampel, 1995), une méthode d'analyse des découvertes technologiques repose sur l'exploitation de ressources en ligne telles que les publications et les bases de données de brevets. Cette méthode vise à approfondir notre compréhension des avancées scientifiques et technologiques en s'appuyant sur l'analyse quantitative via des modèles hybrides qui combinent des méthodes bibliométriques, statistiques, d'exploration de données et de fouille de texte (Porter & Cunningham, 2004). L'objectif est de fournir aux chercheurs, aux décideurs et aux entreprises des informations précieuses sur les évolutions technologiques possibles. L'étude des opportunités et des découvertes d'innovations technologiques s'est enrichie pour inclure non seulement les trajectoires de développement classiques, mais aussi les nouvelles innovations et opportunités technologiques émergentes. Elle examine également les collaborations potentielles et les centres d'innovation en utilisant diverses techniques d’analyse de données (Song *et al*., 2017).

Dans un secteur particulier comme l'industrie du plastique, où le Canada occupe la 6è place mondiale en termes de production de plastiques, avec le Québec se positionnant comme

la deuxième province la plus importante dans l'industrie des composites et des plastiques. Dans la région de Chaudière-Appalaches, cette industrie représente environ 11 % des emplois manufacturiers et génère plus de 1 milliard de dollars de chiffre d'affaires annuellement. À l'échelle mondiale, cette industrie connaît une croissance continue, stimulée par l'utilisation croissante des matériaux composites et des plastiques dans divers secteurs tels que l'automobile, l'électronique, le cosmétique, le médical et la construction. Cependant, les entreprises de la région doivent relever plusieurs défis, notamment la concurrence accrue de la Chine et la guerre des prix. L'innovation, soutenue par la veille technologique, peut grandement aider les entreprises canadiennes et régionales à surmonter ces obstacles et à maintenir leur compétitivité.

De manière traditionnelle, la surveillance technologique dans ce domaine est basée sur l'étude des brevets et des publications scientifiques. Toutefois, avec l'émergence des plateformes numériques et des réseaux sociaux, de nouvelles sources de données sont devenues accessibles, offrant ainsi des perspectives nouvelles et réellement peu exploitées par la veille traditionnelle (Doe, 2019). L'émergence et l'évolution rapide d'innovations technologiques constituent un défi majeur, ce qui nécessite des méthodes de détection robustes.

En intégrant certaines techniques d'apprentissage automatique dans ces procédés, il est possible d'apporter une nouvelle dimension aux techniques de surveillance technologique en utilisant l'analyse automatique de grandes quantités de données textuelles. Cela inclut à la fois les données des brevets, qui offrent des informations précieuses sur les technologies actuelles et les tendances émergentes, ainsi que les échanges en temps réel sur des plateformes telles que Twitter et Web of Science. Les données issues de ces plateformes constituent des sources précoces d'informations pertinentes, permettant d'identifier des innovations émergentes avant leur brevetage ou leur publication officielle (Bessen & Hunt, 2007 ; Gloor, 2017).

Selon Zhang *et al.* (2014) et Porter (2015), des approches hybrides ont été explorées pour extraire des mots et expressions clés à partir de données textuelles, en combinant différentes méthodes pour renforcer la pertinence et la cohérence de l'analyse thématique. Les méthodes basées sur l'analyse sémantique (en anglais *Subject-Action-Object*, SAO) utilisent la structure des phrases pour repérer et suivre les éléments technologiques, offrant ainsi des visualisations dynamiques des évolutions technologiques (Yoon *et al*., 2013 ; Yoon & Kim, 2011 ; Guo *et al*., 2016). Une approche hybride a été proposée par (Chen *et al.*, 2018), qui combine l'analyse de réseau et l'apprentissage automatique pour détecter les nouvelles tendances dans les domaines technologiques. Par ailleurs, Zhang *et al.* (2017) ont proposé une approche hybride qui combine l'analyse sémantique et l'apprentissage profond afin de repérer les avancées dans les secteurs de la biotechnologie et de la médecine. La combinaison de diverses méthodologies a démontré son efficacité pour améliorer l'identification et le suivi des progrès technologiques en cours.

En outre, l'emploi d'applications d'apprentissage automatique a été un véritable succès dans l'anticipation de la révolution et la découverte de nouvelles opportunités de recherche dans des domaines tels que la finance et la santé (Hashimoto *et al*., 2018 ; Kumar & Rahman, 2020). Ces réussites démontrent l'avantage de l'apprentissage automatique pour repérer des signaux faibles et des nouvelles avancées, tout en soulevant d'importantes questions méthodologiques. Quelles mesures peuvent être prises pour garantir que les modèles d'apprentissage automatique saisissent de manière adéquate les nuances des données textuelles ? Quels sont les critères les plus performants pour repérer les avancées récentes dans les données non structurées ? Comment ces instruments peuvent-ils, de manière dynamique et automatique, prévoir non seulement les tendances actuelles, mais également les innovations technologiques à venir ?

Ces interrogations soulignent l'importance de développer une méthode intégrant des techniques avancées de traitement du langage naturel et d'analyse prédictive, spécifiquement adaptées à un domaine d'activité ciblé. L'objectif principal de cette étude est d'élaborer un

cadre méthodologique performant pour l'utilisation de l'apprentissage automatique dans la veille technologique, en mettant l'accent sur la détection des signaux précurseurs d'innovations radicales. Cette étude s'appuie sur une méthodologie à la fois éprouvée et innovante pour identifier les signaux avant-coureurs des nouvelles innovations dans le secteur du plastique. Elle fait appel à des techniques avancées de traitement du langage naturel, telles que le calcul de similarité des mots-clés technologiques et l'analyse prédictive, adaptées au contexte spécifique de cette industrie.

1.1.2. Problématique

La veille technologique est un processus clé pour maintenir la compétitivité des entreprises dans divers secteurs, en les aidant à anticiper les avancées technologiques et les évolutions du marché. L'identification des opportunités technologiques, c'est-à-dire la capacité à repérer, évaluer et exploiter de nouvelles idées, repose souvent sur des analyses de données issues de brevets et de publications scientifiques (Song *et al.*, 2017 ; Park & Yoon, 2018). Les méthodes traditionnelles, basées sur ces sources statiques, se révèlent toutefois limitées par l'absence de données en temps réel, notamment celles provenant des réseaux sociaux, qui jouent un rôle croissant dans la détection précoce d'innovations (Doe, 2019).

Les méthodes traditionnelles, bien que robustes, reposent souvent sur une intervention manuelle et des approches telles que l'analyse de la fréquence des mots-clés, ce qui peut générer des résultats biaisés ou incomplets (Yoon & Kim, 2012). L'intégration de techniques d'apprentissage automatique et d'analyse de réseaux, en particulier pour le traitement de données non structurées telles que des textes de brevets ou des publications sur les réseaux sociaux, offre de nouvelles perspectives pour identifier des technologies émergentes et inexploitées (Lee *et al.*, 2020). L'essor des algorithmes de traitement automatique du langage naturel (NLP) permet également une analyse plus fine des documents, en identifiant des expressions spécifiques révélatrices d'innovations (Green *et al.*, 2021).

Par ailleurs, l'utilisation de techniques avancées telles que la modélisation thématique et l'analyse de similarité des textes permet de détecter des tendances technologiques dans de grandes quantités de données, offrant ainsi une vision plus complète d'opportunités technologiques (Kousis & Tjortjis, 2023). L'intégration de sources de données en temps réel, telles que les réseaux sociaux aux bases de données traditionnelles comme les brevets, permet d'anticiper plus efficacement les tendances émergentes et d'optimiser le processus de veille technologique. Cette exploitation d'informations offre aux entreprises la capacité de détecter rapidement des innovations potentielles tout en ajustant leur stratégie de recherche et développement en fonction des dynamiques du marché. Une telle approche favorise ainsi une identification proactive et efficace des opportunités technologiques novatrices.

Cette étude propose une approche automatisée et dynamique, visant à minimiser l'intervention humaine par une combinaison des sources traditionnelles et des flux de données numériques en temps réel. Cette approche permet de détecter et de suivre les tendances technologiques émergentes dans l'industrie du plastique, facilitant ainsi l'identification des opportunités d'innovation et le suivi des dynamiques industrielles en temps réel. Ce secteur, particulièrement compétitif à l'échelle mondiale, nécessite une innovation constante pour rester performant, notamment face aux pressions des pays à faible coût de production (Brown *et al*., 2022). En appliquant des techniques de traitement automatique du langage naturel et d'apprentissage automatique, cette recherche aspire à fournir aux entreprises des outils pour mieux anticiper les changements technologiques et ajuster leurs stratégies de recherche et développement en conséquence (Liu *et al*., 2024).

1.2. OBJECTIFS ET APPROCHES MÉTHODOLOGIQUES

1.2.1. Objectifs de la recherche

Pour mesurer l'efficacité de cette méthode, nous établirons des indicateurs clés, tels que le taux de similarité entre les innovations identifiées et celles déjà brevetées. Ce taux servira à évaluer la capacité de la méthode à repérer des innovations en phase avec les développements brevetés, offrant ainsi une mesure de la pertinence et de la précision des résultats. Cette approche sera appliquée spécifiquement au secteur du plastique, où l'innovation est essentielle pour relever les défis de durabilité et répondre aux exigences croissantes en matière de développement durable. Pour atteindre cet objectif, nous exploitons des techniques sophistiquées de traitement du langage naturel et d'apprentissage automatique. Ces techniques permettent de traiter de grandes quantités de données textuelles provenant des sources non structurées, telles que les réseaux sociaux et de les comparer avec des données structurées, telles que celles contenues dans les brevets de l'Office américain des brevets et des marques (USPTO) (Black & Green, 2020). L'utilisation de ces outils permet d'automatiser l'extraction des données et d'effectuer des analyses de similarité textuelle, ce qui facilite la détection précoce d'innovations potentielles et l'évaluation de leur originalité par rapport aux technologies existantes.

En appliquant cette méthodologie à l'industrie du plastique, nous visons à identifier des opportunités technologiques telles que les nouveaux matériaux, les méthodes de fabrication durables et les technologies de recyclage innovantes. Ce choix de secteur est motivé par l'importance économique du plastique et les défis spécifiques auxquels l'industrie est confrontée, notamment la nécessité de réduire son impact environnemental et de répondre à une concurrence accrue. En intégrant les données en temps réel des réseaux sociaux avec les informations issues des brevets, cette étude permet de développer une approche dynamique de la veille technologique. Cette approche aide non seulement à identifier les tendances émergentes, mais aussi à anticiper les évolutions futures dans un secteur en constante mutation. Les résultats de cette recherche fourniront des informations stratégiques

cruciales pour les entreprises et les décideurs, telles que des prévisions sur les tendances technologiques à venir, leur permettant de prendre des décisions éclairées et de rester compétitifs dans un environnement technologique en rapide évolution.

1.2.2. Approche méthodologique

Dans cette étude, nous proposons une approche stratégique intégrant diverses disciplines et techniques pour aborder la veille technologique de manière automatique et innovante, avec pour objectif de contribuer de manière significative à la compréhension des dynamiques d'innovation dans le secteur du plastique (voir Figure 1). La méthode employée repose sur plusieurs aspects technologiques et analytiques :

Collecte des données : Nous recueillons les informations provenant de différentes sources afin d'obtenir un ensemble d'informations variées. Ce recueil comprend des tweets, des résumés scientifiques et des informations de brevets. Pour extraire des informations en temps réel sur les discussions technologiques, des outils de collecte de données sur le web et des APIs tels que *ntscraper* sont employés pour les *tweets*. Les résumés scientifiques sont obtenus en utilisant des bases de données académiques comme Web of Science. En accédant aux archives de la base USPTO, la base de données américaine de brevets, on peut collecter des brevets.

Prétraitement des données : Les informations recueillies sont soumises à un prétraitement pour assurer leur qualité et leur cohérence. Cela implique la standardisation des textes, l'élimination des caractères spéciaux et la suppression des doublons. La division des textes en unités significatives (tokens) est effectuée par la tokenisation, tandis que la lemmatisation réduit les mots à la forme initiale. Il est essentiel de passer par cette étape afin de préparer les données à l'analyse et de réduire les biais causés par des incohérences ou des erreurs dans les données brutes.

Calcul de similarité : Afin d'évaluer la similarité entre les documents textuels, nous employons des modèles de traitement du langage naturel (NLP) avancés tels que BERT

(Bidirectional Encoder Representations from Transformers) et RoBERTa (Robustly Optimized BERT Pretraining Approach). Ces modèles sont choisis pour leur capacité à capturer les nuances sémantiques et contextuelles, en convertissant les textes en vecteurs denses. En appliquant la mesure de similarité cosinus, nous pouvons identifier les documents ayant des thèmes similaires et repérer des innovations émergentes par rapport aux brevets existants. BERT et RoBERTa offrent une performance avérée et une flexibilité qui enrichissent notre analyse, permettant ainsi une meilleure compréhension des dynamiques technologiques dans le secteur du plastique.

Comparaison avec les archives de brevets de l'USPTO : La comparaison des tendances et des sujets identifiés avec les archives de brevets de l'USPTO permet de confirmer les tendances émergentes en vérifiant leur inscription dans les bases de données de brevets existantes. Le but est de valider des innovations repérées ou identifiées et de vérifier qu'elles correspondent à des avancées technologiques déjà brevetées ou en cours de brevetage.

Utilisation de modèles d'apprentissage automatique : L'utilisation des modèles d'apprentissage automatique permet de trier et de regrouper les données textuelles afin d'approfondir l'analyse. Ces modèles facilitent l'identification de nouvelles tendances en analysant de vastes quantités de données et en découvrant des relations complexes entre les différentes sources d'information.

Clustering et découverte de sujets latents : Le calcul de la similarité et l'utilisation de la méthode d'allocation latente de Dirichlet (LDA) permettent de décomposer les documents en sujets latents. La méthode d'allocation latente de Dirichlet est un algorithme qui permet de repérer des groupes de mots souvent associés dans les textes. Cela permet de repérer des thèmes sous-jacents dans de vastes corpus de données. Cette étape permet de classer les données en catégories importantes pour faciliter l'identification des tendances et des domaines technologiques en plein essor.

Cette approche intégrée des données basée sur une méthode automatique permet de détecter précocement les avancées et les tendances technologiques dans le secteur du plastique. Notre capacité à identifier des signaux précoces et à anticiper les évolutions futures

du marché est renforcée par la combinaison de données provenant des réseaux sociaux et des brevets traditionnels.

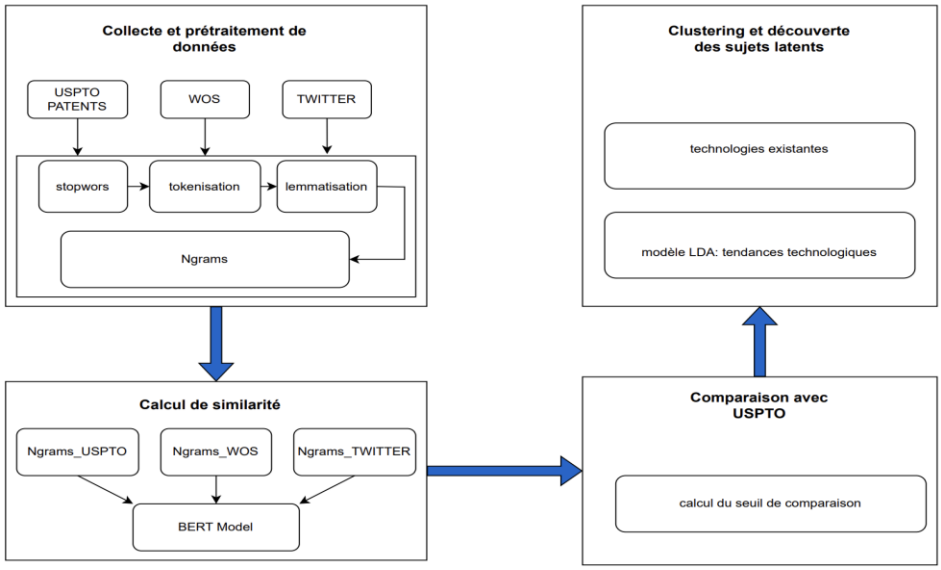

Figure 1. Approche méthodologique.

1.3. CONTRIBUTIONS

Cette étude propose une approche innovante pour renforcer la surveillance technologique, en intégrant des données non conventionnelles issues des réseaux sociaux et des résumés académiques, en plus des brevets. En utilisant des techniques avancées de traitement du langage naturel (NLP) et d'apprentissage automatique, cette étude permet de détecter des tendances émergentes et des innovations non brevetées qui échappent aux méthodes traditionnelles. Cette approche offre une vision claire des opportunités technologiques et des obstacles potentiels à leur adoption, notamment dans le secteur du plastique.

L’originalité de cette approche réside dans la recherche proactive des sujets technologiques non encore définis dans les brevets existants, permettant ainsi de cartographier les horizons inexplorés. Grâce à l’application de techniques statistiques, l’étude valide les résultats en comparant systématiquement les données provenant de sources alternatives avec celles des brevets, identifiant des domaines disruptifs. Cette approche permettra aux entreprises de repérer les signaux faibles et d’orienter plus efficacement leurs investissements en recherche, développement et innovation, favorisant ainsi une prise d’avance sur les évolutions technologiques. Cette capacité d’anticipation constitue un avantage stratégique majeur, permettant aux entreprises de se positionner à la pointe de l'innovation et d’adopter des pratiques durables.

1.4. STRUCTURE DU MÉMOIRE

Ce mémoire par articles est structuré en quatre chapitres principaux : introduction, état de l'art, application de l'apprentissage automatique à la veille technologique et conclusion. Le premier chapitre, introduction, présente le contexte de l'étude, les objectifs de recherche ainsi que les principaux concepts dans le domaine de la surveillance technologique et de l'innovation dans l'industrie des plastiques. Le deuxième chapitre, l’état de l'art, passe en revue les recherches et les connaissances existantes, analyse les méthodes actuelles de surveillance technologique et identifie les lacunes dans la littérature. Le troisième chapitre, application de l'apprentissage automatique à la veille technologique, présente l'article scientifique de l'étude, détaillant les méthodes utilisées, les analyses réalisées, les résultats obtenus et les implications pour la veille technologique et l'innovation dans le secteur du plastique. Le quatrième chapitre, conclusion, met en évidence les contributions aux connaissances et à la pratique et propose des recommandations pour les recherches futures et les applications pratiques. En complément, une bibliographie répertorie toutes les sources et références citées, tandis que le second article est présenté en annexe.

CHAPITRE 2 : ÉTAT DE L'ART

2.1. INTRODUCTION

Le secteur du plastique au Canada génère environ 25 milliards de dollars en revenus annuels et emploie plus de 100 000 personnes, principalement en Ontario et au Québec, qui assurent plus de 80 % de la production nationale (Petigny, Ménigault *et al*., 2019). Ce secteur a transformé des domaines tels que l'emballage et l'automobile en développant des produits innovants qui améliorent la compétitivité économique des entreprises (Gonçalves, Cardeal *et al.*, 2024). Cependant, cette croissance rapide a engendré des défis environnementaux majeurs, tels que la pollution par les microplastiques et la contamination marine. Pour surmonter ces défis, il est crucial d'explorer et de découvrir de nouvelles technologies pour traiter et optimiser l'utilisation du plastique au Canada, notamment en adoptant des techniques d'apprentissage automatique pour identifier les innovations et les opportunités technologiques.

Les chercheurs ont historiquement utilisé des approches qualitatives pour l'analyse de texte, telles que le codage manuel, l'analyse du discours et la théorie ancrée (Duriau *et al*., 2007). Cependant, ces méthodes ont montré leurs limites face à des volumes croissants de données textuelles (Kobayashi *et al*., 2018). Ainsi, l'analyse assistée par ordinateur, notamment la fouille de texte, s'est imposée comme une solution indispensable pour traiter efficacement de grands corpus textuels de manière transparente et reproductible (Wiedemann, 2013).

Cette étude propose de développer une approche dynamique pour la découverte d'opportunités technologiques en utilisant des techniques de fouille de texte basée sur l'assistance par ordinateur. L'objectif est d'identifier et de classer les nouvelles technologies en analysant des données issues d'articles scientifiques, de tweets technologiques et de brevets. Cette approche utilise des modèles de traitement automatique du langage naturel (TAL) et des méthodes de modélisation des sujets.

2.2. VEILLE TECHNOLOGIQUE

La veille technologique est une approche méthodique visant à recueillir, analyser et diffuser des données sur les avancées technologiques et les tendances émergentes. Selon la norme UNE 166006 :2011, qui régit la gestion de la recherche, du développement et de l'innovation (RDI), il s'agit d'un processus structuré et continu permettant de collecter des informations provenant à la fois de l'extérieur et de l'intérieur de l'organisation, dans les domaines scientifique et technologique. L'objectif est de sélectionner, d'étudier, de diffuser et de partager ces informations afin de les transformer en connaissances exploitables, favorisant ainsi une prise de décision éclairée et une meilleure anticipation des évolutions. La veille technologique joue un rôle central dans le processus de RDI en réduisant les risques liés aux choix stratégiques et en renforçant la compétitivité de l'organisation (Perez *et al*., 2018). (Voir Figure 2).

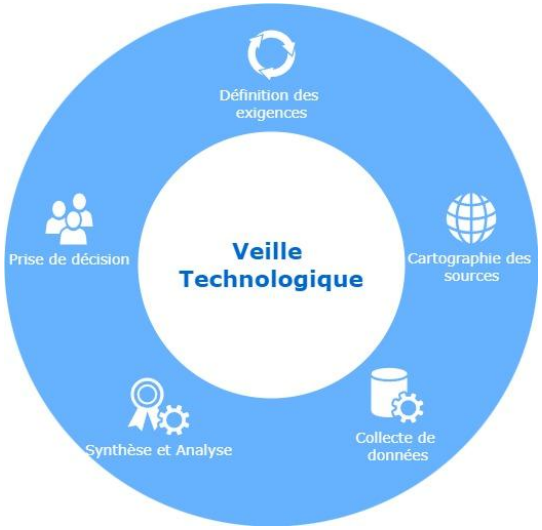

Figure 2. Processus de veille technologique.

Le processus de veille technologique s'articule autour de cinq étapes essentielles (Kahaner, 1997), visant à définir les exigences, cartographier les sources, collecter, analyser et diffuser des informations stratégiques sur les avancées technologiques et les tendances émergentes.

La première étape consiste à définir des exigences, qui établissent les objectifs spécifiques de la veille, tels que la détection d'innovations ou l'anticipation des évolutions technologiques. Cette phase permet également de définir les axes thématiques à explorer et les types d'informations nécessaires, assurant ainsi une direction claire et ciblée pour l'ensemble du processus.

La deuxième étape est la cartographie des sources, qui structure la veille en identifiant et sélectionnant des sources d'information pertinentes et fiables. Ces sources incluent les bases de données de brevets, les publications scientifiques, les rapports de recherche ainsi que les réseaux sociaux et forums spécialisés. Bien que cette phase puisse être chronophage, elle est cruciale pour garantir la qualité et la pertinence des données à collecter.

Une fois les sources définies, la collecte d'informations constitue la troisième étape. Elle s'effectue via une variété de canaux, combinant des outils traditionnels et numériques. L'utilisation de systèmes de surveillance automatisée, tels que des alertes ou des flux RSS, permet une collecte continue et en temps réel, capturant ainsi les dernières évolutions technologiques.

La quatrième étape, l'analyse des données collectées, est une phase critique du processus de veille. En exploitant des techniques avancées de traitement automatique du langage naturel (NLP) et d'apprentissage automatique (ML), cette phase permet de transformer de grands volumes de données en connaissances exploitables. Selon Chau Minh & Reuter (2024), ces méthodologies facilitent l'identification des tendances émergentes, la détection d'innovations clés et l'établissement de relations significatives au sein des données grâce à des modèles thématiques et des mesures de similarité.

Enfin, la cinquième étape est la diffusion des résultats, qui consiste à partager les informations extraites sous forme de rapports détaillés, de tableaux de bord interactifs ou d'alertes stratégiques. Cette diffusion, ciblant les parties prenantes pertinentes, joue un rôle fondamental en orientant la prise de décision et en alignant les actions avec les objectifs organisationnels.

Ces étapes intégrées s'appuient sur des disciplines telles que la recherche et le développement (R&D), l'intelligence économique et les technologies de l'information et de la communication (TIC). Historiquement, elles ont été déterminantes pour maintenir la compétitivité des entreprises et stimuler l'innovation, un rôle qui s'est renforcé grâce aux avancées constantes dans le domaine des technologies de l'information.

2.3. DÉCOUVERTE D'INNOVATIONS TECHNOLOGIQUES

L'essor de l'apprentissage automatique et du traitement du langage naturel (NLP) a révolutionné la capacité à traiter et analyser de grandes quantités de données, ouvrant de nouvelles perspectives pour la veille technologique (Perez et al., 2018). Ces avancées permettent de surmonter les limites de l'analyse manuelle en offrant des solutions innovantes pour gérer des ensembles de données complexes et volumineux (Armentano et al., 2014). En effet, les algorithmes d'apprentissage automatique, combinés aux techniques avancées de NLP, s'avèrent particulièrement efficaces pour extraire des informations pertinentes à partir de vastes corpus textuels.

Les opportunités technologiques, essentielles pour les organisations cherchant à innover, se trouvent souvent dispersées dans divers types de données, telles que les articles scientifiques, les brevets et les rapports de recherche (voir Figure 3). Pour identifier et exploiter ces opportunités efficacement, des méthodes et outils scientifiques adaptés sont nécessaires. La détection de ces opportunités au sein de vastes volumes de données devient ainsi un sujet de recherche crucial et complexe (Wang *et al.*, 2023).

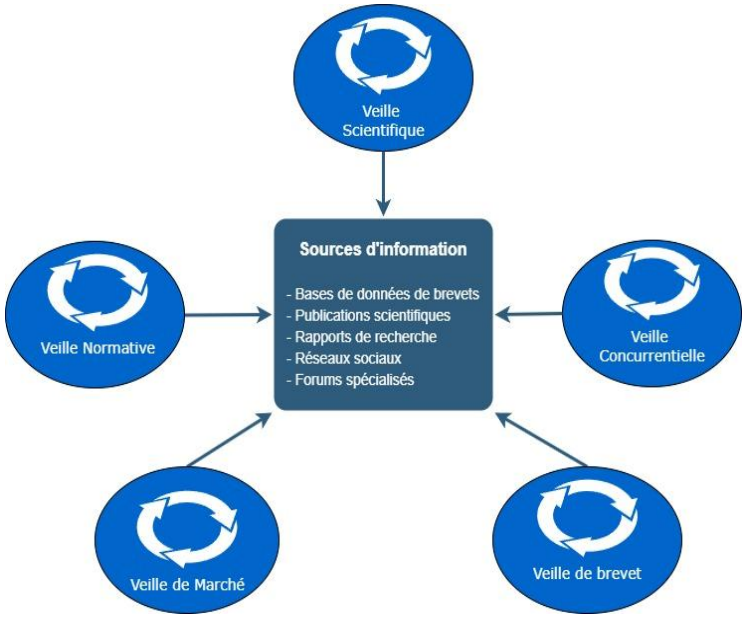

Figure 3. Sources d'information et types de veille.

Cependant, malgré les avancées offertes par l'apprentissage automatique et le traitement du langage naturel, plusieurs défis et limitations demeurent dans les approches existantes. L'hétérogénéité des données provenant de sources variées complique leur intégration, ce qui peut entraîner une analyse fragmentée et restreindre la capacité à tirer des conclusions robustes sur les tendances technologiques (Wang *et al*., 2023). Les techniques de fouille de texte, bien que efficaces pour identifier des motifs, souffrent souvent d'une interprétation limitée des contextes, ce qui peut mener à des résultats superficiels. Par exemple, les méthodes telles que TF-IDF et LDA, tout en étant utiles pour la modélisation thématique, négligent souvent les relations sémantiques profondes entre les termes, limitant leur efficacité à détecter les tendances émergentes dans un paysage technologique dynamique (San *et al.*, 2021). De plus, l'utilisation de modèles comme les machines à vecteurs de support (SVM) et les arbres de décision peut entraîner des biais en raison de leur dépendance à des caractéristiques prédéfinies, ce qui peut fausser l'identification des opportunités technologiques (Perez et al., 2018). Bien que des approches, telles que celles proposées par Zhu et Porter (2002), utilisent des outils comme VantagePoint pour automatiser partiellement la classification des informations pertinentes, elles peinent à s'adapter aux évolutions dynamiques des données technologiques, limitant ainsi leur flexibilité.

Dans cette optique, Wang *et al*., (2015) ont développé une feuille de route technologique fondée sur l'approche Sujet-Objet-Action (SOA), conçue pour analyser les tendances technologiques en associant les termes techniques à leurs fonctions spécifiques. Bien que cette méthode offre une structure utile, elle requiert des ressources considérables et une intervention humaine substantielle, restreignant son efficacité dans les applications de veille technologique automatisée. De plus, Perez *et al.* (2018) ont intégré des techniques de fouille de texte et d'apprentissage automatique pour créer un moteur de classification capable de filtrer automatiquement les informations d'un système de veille technologique, réduisant ainsi partiellement le temps de travail des annotateurs humains. Toutefois, cette approche ne supprime pas entièrement le besoin d'intervention humaine, ce qui souligne la nécessité d'optimiser encore ces méthodes. En outre, Zhou *et al.* (2020) ont utilisé une méthodologie scientométrique pour identifier les mots-clés technologiques les plus pertinents, analysant le volume de publications et de brevets dans des domaines spécifiques. Cette étude a permis de générer des prévisions d'innovations technologiques. Cependant, elle n'a pas intégré les caractéristiques sémantiques internes des brevets, ce qui peut biaiser l'interprétation des résultats liés à l'innovation

technologique. Par ailleurs, d'autres sources d'information, telles que les réseaux sociaux et les articles scientifiques, constituent des réservoirs précieux pour identifier les opportunités technologiques. Ces sources méritent une exploration approfondie grâce à des techniques d'apprentissage automatique, permettant de découvrir des modèles et des tendances significatifs. Par exemple, Hu *et al.* (2024) ont réussi à combiner des données d'articles scientifiques et de brevets pour analyser les technologies émergentes. Ils ont utilisé des méthodes telles que la co-citation et le couplage bibliographique pour mesurer la similitude thématique entre les publications scientifiques.

L''intégration des modèles LDA et BERT s'avère prometteuse pour surmonter ces limitations. Elle permet d'améliorer la capture des relations sémantiques grâce à des représentations contextuelles riches et d'affiner l'identification des sujets et des tendances. Cette combinaison favorise l'établissement de liens entre des termes qui pourraient sembler déconnectés dans une analyse traditionnelle. L'utilisation du modèle BERT pour l'extraction d'informations et la classification confère une flexibilité considérable, rendant possible une adaptation rapide aux évolutions technologiques. Cette capacité à anticiper proactivement les opportunités inexploitées renforce la précision des analyses, exploitant pleinement la richesse des données textuelles. Cette méthodologie surmonte les limites des approches antérieures basées sur des données historiques ou des modèles rigides. Ainsi, notre recherche se positionne pour offrir des perspectives précieuses, contribuant à une veille technologique plus dynamique et réactive, en intégrant des approches innovantes et l'utilisation combinée d'algorithmes performants pour surmonter les limites des méthodes traditionnelles.

La Table 1 présente un panorama des méthodes principales employées dans le cadre de la veille technologique, en les associant aux auteurs de référence, à une description de chaque méthode, et aux domaines d'application spécifiques dans lesquels elles sont déployées. Ces méthodes, issues de différentes disciplines et approches analytiques, permettent de traiter, d'analyser et de classer de grandes quantités de données pour en extraire des informations exploitables, soutenant ainsi l'innovation et le développement stratégique dans divers secteurs technologiques.

Table 1. Synthèse des méthodes et approches pour la veille technologique.

Méthodes	**Références/auteurs**	**Descriptions de méthodes**	**Domaines d'application**
Carte de route technologique (TRM)	Wang *et al*., 2015.	Sujet-Action-Objet (SAO)	Technologie photovoltaïque
Avis d'expert	Perez *et al.*, 2018.	Évaluation qualitative, Keyword Cluster, filtrage des sujets et construction de moteur de filtrage automatique	Technologie
Fouille de texte	Zhu & Porter (2002).	Analyse d'opportunité technologique (TOA)	Intelligence technologique
Modélisation thématique	Daniel San *et al*. (2021).	Surveillance Automatique technologique	Web et réseaux sociaux
Algorithmes de regroupement et sac des mots	Yuan Zhou *et al*. (2020).	Clustering en courbe de S, approche scientométrique	Apprentissage profond

2.4. CONCLUSION

L'intégration des systèmes de veille technologique avec les outils d'apprentissage automatique et de traitement automatique du langage naturel permet non seulement de suivre les trajectoires technologiques actuelles, mais aussi de prédire les futures percées avant qu'elles ne s'imposent dans le grand public. Cette évolution ouvre de nouvelles perspectives pour la découverte d'innovations radicales, révolutionnant ainsi les méthodes traditionnelles de R&D et offrant une compétitivité accrue aux entreprises.

Cette étude a pour objectif principal de développer une méthodologie dynamique qui tire automatiquement et pleinement parti des flux de données en temps réel des plateformes numériques et des réseaux sociaux, en complément des sources traditionnelles telles que les brevets, pour identifier et suivre les tendances technologiques émergentes. Cette approche vise à fournir des aperçus plus précis et opportuns, permettant aux entreprises de mieux anticiper les évolutions technologiques et de prendre des décisions stratégiques plus éclairées en matière de R&D.

CHAPITRE 3 : APPLICATION DE L'APPRENTISSAGE AUTOMATIQUE À LA VEILLE TECHNOLOGIQUE

3.1. RÉSUMÉ EN FRANÇAIS DU PREMIER ARTICLE

L'industrie du plastique est cruciale pour l'économie du Canada, en particulier au Québec. Les défis environnementaux persistent et les entreprises investissent dans la recherche pour améliorer les performances et la durabilité des produits. Les développements récents comprennent les polymères biodégradables et les matériaux composites. Cette recherche vise à élaborer une méthode automatisée d'extraction et d'analyse des données textuelles à l'aide d'une analyse de la similitude de texte et de la modélisation du sujet LDA. Ce processus identifie les innovations brevetées existantes et les nouvelles, créant des catégories supplémentaires au sein du système de classification des brevets. Le modèle ROBERTa utilisé par BERT, formé sur les données sur les brevets, rend plus efficace l'identification de la similitude sémantique entre les classes technologiques et leurs résumés de brevets avec une précision nettement supérieure à 80 %, quel que soit le seuil de similitude établi. L'analyse du sujet LDA a établi un score de 52 % de cohérence du sujet. L'examen des résumés des publications académiques de Web of Science a révélé, par exemple, des approches transitoires à l'économie circulaire, qui représentent une autre option viable pour gérer la fin de vie des plastiques tout en réduisant la pollution de l'environnement.

Cet article, intitulé « *Application of machine learning in technological forecasting* », a été soumis et accepté pour présentation à la conférence EUSPN 2024, *The 15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks*, qui s'est tenue du 28 au 30 octobre 2024 à Leuven, Belgique (Tshibanda *et al.*, 2024). En tant que premier auteur, j'ai contribué principalement à la recherche sur l'état de la question, au développement de la méthodologie et à son opérationnalisation. Adda Mehdi, second auteur, a aidé à la recherche sur l'état de la question, au développement de la méthode ainsi qu'à la révision de l'article. Said Echchakoui, le troisième auteur, a fourni l'idée originale, a aidé à la recherche sur l'état de la question et a également contribué à la revue de la littérature.

3.2. APPLICATION OF MACHINE LEARNING IN TECHNOLOGICAL FORECASTING

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Application of machine learning in technological forecasting

Franck Tshibanda Nkolongo[a,*], Adda Mehdi[b], Said Echchakoui[a,b]

[a]*Department of Mathematics, Computer Science and Engineering, University of Quebec at Rimouski, Canada*
[b]*Management Sciences Departmental Unit, University of Quebec at Rimouski (Lévis), Canada*

Abstract

The plastics industry is crucial to Canada's economy, particularly in Quebec. Environmental challenges persist, and companies invest in research to improve product performance and sustainability. Recent developments include biodegradable polymers and composite materials. This research aims to develop an automated method for extracting and analyzing text data using text similarity analysis and LDA subject modeling. This process identifies existing patented innovations and new ones, creating additional categories within the patent classification system. The ROBERTa model used by BERT, trained on patent data, makes it more effective to identify semantic similarity between technological classes and their patent summaries with an accuracy significantly greater than 80%, regardless of the similarity threshold established. The LDA subject analysis established a 52% subject consistency score. The review of Web of Science's academic publication summaries revealed, for example, transitional approaches to the circular economy, which represent another viable option for managing the end-of-life of plastics while reducing environmental pollution.

Keywords: Patent analysis; Plastic industry; BERT; RoBERTa; LDA; Technological opportunities; Innovation.
Type your keywords here, separated by semicolons ;

1. INTRODUCTION

The plastic sector in Canada, which generates 25 billion dollars annually [1], has transformed industries such as packaging and automobiles, increasing economic competitiveness [2]. However, rapid growth has led to increased plastic waste, environmental issues like microplastic pollution and plastic landfills. To address these issues, new methods and technologies must be developed to treat plastics and improve their daily use in Canadian society. A practical approach is using machine learning to identify technological advances and opportunities in this area. Technological exploration, first defined by [3], uses text mining tools to extract scientific, technological, and innovative information. Its aim is to generate practical insight for decision-making in technological vigilance, process management, and science and technology indicators. Recent advances have solidified its importance in innovation management [4].

Researchers have analyzed trends in technological discovery. [13] proposed a co-citation and bibliographic coupling analysis to explore the field of technological discovery, thus revealing its foundations and evolution. For example, [5] summarized the evolution of bibliometry, text analysis, and visualization to extract relevant information. However, these manual methods are extremely laborious and show limits in the face of growing volumes of text data [6]. In fact, the larger the body of text, the more it becomes necessary to use automated techniques, while manual encoding remains relevant for smaller datasets. To overcome these constraints, computer-assisted analysis of textual data was explored.

This study uses a dynamic approach to discover technological opportunities in the plastics industry by extracting text from various data sources. It uses 31 articles from Clarivate's Web of Science bibliographic database, 47 recent technology tweets, and an extensive patent database to classify new technologies. The study aims to develop advanced research using automated natural language processing, using BERT models for text similarity calculations, and LDA subject modeling to capture and classify innovations. This approach can help clarify complex issues in the plastics industry.

The remainder of this article is structured as follows. Section 2 gives an overview of the relevant literature; Section 3 details the research methods and data used; Section 4 presents the results related to models at the subject cluster level. It also discusses the implications of these findings; section 5 concludes the article, summarizing key points and suggesting directions for future research.

2. Literature Review

In this study, we explore the impact of natural language processing and machine learning on marketing in the plastics industry. We highlight the potential of Big Data to discern technical and economic models while identifying three main obstacles: access to large amounts of data, their efficient management, and the development of the necessary technological skills. Different types of data, such as scientific papers, patents, and research reports, contain dispersed technological possibilities. The effective identification and exploitation of these opportunities requires scientific methods, thus constituting a complex and crucial area of research. [7]. Innovation specialists have developed technology forecasting and roadmap tools based on patent analysis [8].

Patents play an essential role in evaluating research progress, and we use text extraction methods to discover patterns in large text collections [11]. One of the main challenges of knowledge management systems is the effective discovery and exploitation of the content of knowledge bases. [9]. Recent progress in text mining has strengthened the importance of technological exploration in innovation management [4]. We use techniques such as natural language processing and data analysis to optimize the exploitation of patent information [10]. For example, [12] used network and thematic analyses to evaluate AI technologies of semiconductor manufacturers. [11] examine the publication and distribution of patents, emerging technologies, and industrial promotion policies in China and the United States. [13] used co-citation analysis and bibliographic coupling to identify current and future research trends.

In addition to patents, we view social networks and scientific articles as valuable sources of information on technological possibilities, using machine learning techniques to identify patterns and trends in these data [15].

Previous research on identifying technological opportunities has several shortcomings, including focusing mainly on scientific papers and patents, requiring manual intervention, and being subjective. This study proposes an approach to automatically identify patented technologies and new opportunities in structured and unstructured data. By integrating data sources and automating data extraction, this method offers a proactive approach to the plastic industry. It uses text similarity analysis and LDA thematic modelling techniques to provide valuable insights into trends and innovations, helping to maintain a competitive advantage in an ever-changing market.

3. Methodology

Our automated technology forecasting methodology for the plastics industry integrates data extraction, identifies current technologies Fig. 1. For this study, we adopted an integrated methodological approach to comprehensively analyze data from multiple sources, such as patent databases, social media platforms for tweets, and scientific databases for article abstracts. The initial data collection relied on APIs to ensure structured and complete retrieval of relevant information.

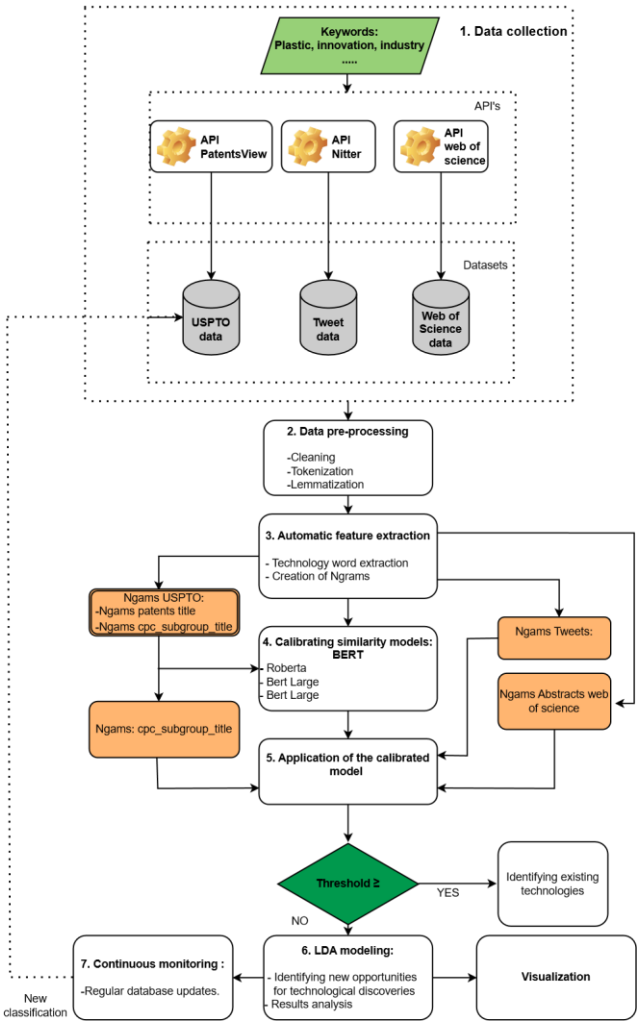

Fig. 1. Methodological process

The collected data underwent rigorous preprocessing, including information normalization and the application of advanced natural language processing techniques such as n-gram extraction and the use of pre-trained models such as RoBERTa [18]. These steps were critical in improving the semantic representation of the texts and in preparing the data for similarity analysis and topic modeling. In addition, the study used the Word2Vec architecture [17] to analyze both the syntax and semantics of word relationships, identify complex links, and pinpoint technological keyword associations for effective monitoring.

To evaluate the similarity between patents, tweets, and academic abstracts, we developed and validated several semantic similarity models, including distilbert-base-uncased, bert-base-nli-mean-tokens, and RoBERTa-base. Each model was trained on large corpora of data to capture the semantic nuances and relationships between the technical concepts present in the texts [18].

The validation of these models was carried out using standard measures such as cosine similarity, as described in formula 1, which allows the semantic proximity of phrase pairs to be assessed [14]. A rigorous process of comparing the performance of each model on annotated test data sets was used to select the optimal model, RoBERTa, which was recognized for its exceptional ability to accurately and consistently identify relevant semantic relationships.

$$CosineSimilarity = \frac{\vec{A}\cdot\vec{B}}{\|A\|\cdot\|B\|} \quad (1)$$

Where vectors $\|A\|$ and $\|B\|$ are embeddings. The cosine similarity ranges between -1 and 1: 1 indicates that the vectors are identical (very similar phrases), 0 indicates that they are orthogonal (no similarity), and -1 indicates that the vectors are opposite.

The RoBERTa model is used to classify patented technologies on the basis of their relevance and future innovation potential, providing a comprehensive understanding of technological dynamics in the plastics industrial sector.

First, these integration models are applied to USPTO patent data to calibrate, evaluate, and validate the model as the data is annotated. Similarities between the text of patent abstracts and their respective classifications (CPC subgroup titles) will be detected and analyzed. Subsequently, the approved model is implemented on information from technology tweets and academic article abstracts (Web of Science), with the aim of identifying, based on thresholds, hot patented technologies and current opportunities.

This methodology enables effective monitoring of technological developments in the plastic industry, taking advantage of the latest advances in machine processing of natural language and machine learning.

4. Results and Discussion

We present the results of the methodology described in Section 3. The USPTO patent database identified 1,052 different technology classes in the plastic industry, each corresponding to a specific innovation. These 1,052 classes are found in 96 patents, indicating that patents can group multiple technological classes, highlighting diverse advances in the plastics industry.

After preprocessing the data with Gensim's standard simple pre-process tokenizer, we tokenized the data, as shown in Fig. 2 of the word clouds. Each word was represented as vectors in a continuous semantic space, facilitating the extraction of important N-grams from USPTO patent summaries, patent classes, academic articles from the Web of Science (WoS) database, and tweets about advances in the plastic industry.

By extracting N-grams, we identified significant linguistic patterns in our dataset. Our methodology, implemented in four different corpora (patent classes and abstracts, academic articles, and tweets), demonstrated its ability to identify relevant and informative text patterns such as *"curing plastic"*., *"welding plastic"*, *"plastic recycling"*, *"plastic waste"*, and *"fiber reinforce plastic"*

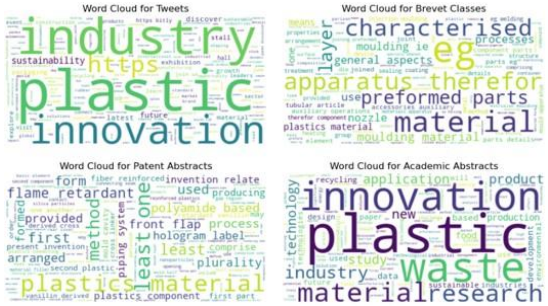

Fig. 2. word clouds

The table 1 illustrates the data sources used for this study, the APIs for automatic data collection, but also the combinations of technology characteristics or keywords in Ngrams regular expressions.

Table 1. Datasets and features.

Data sources	API	Technology word Ngrams
USPTO	PatentsView	ngrams classes, ngrams patent abstract
Twitter	Nitter	ngrams tweets
Web of Science	web of science	ngrams abstracts

4.1. Semantic Similarity

• Validation of the Similarity Model

By carefully analyzing the semantic similarity between 1052 patent summaries and their corresponding 1052 classes, we confirmed our model by performing a total of 1 107 504 similarity pair calculations.
After evaluating the quality of semantic similarity models using three popular BERT variants—distilbert-baseuncased, bert-base-nli-mean-tokens, and RoBERTa-base—we obtained the following results. Each model was trained on a large corpus to capture semantic subtleties and relationships between words in the text. Among these, RoBERTabase emerged as the most effective model, demonstrating superior performance in terms of both accuracy and robustness in capturing semantic similarities.

The models were evaluated through several rigorous steps. Initially, we used cosine similarity to assess the semantic proximity between pairs of phrases. This measure compares the semantic representation vectors of phrases according to the BERT models by calculating the angle between these vectors.

Then, we constructed a similarity matching table by comparing each model's results with a set of annotated test data, as illustrated in Fig. 3. This approach allowed us to objectively quantify and compare the performance of each model in terms of accuracy and its ability to identify relevant semantic relationships in the texts analyzed.

The roBERTa (Robustly Optimized BERT Approach), an improved version of BERT with an optimized architecture and intensive drive, has emerged as the best performance among the evaluated models. It has demonstrated a high ability to grasp semantic subtleties and generate consistent and accurate similarity scores, outperforming other models in most usage situations. According to Fig. 4, the roBERTa model obtained a similarity score of 85.42% from a threshold of 80%, while the bert-large-nli-mean-tokens model received 15.62% and Allenai Scibert Scivocab
14.58%. This threshold will be used as a starting point for identifying new technological possibilities.

	Patent_Classe	Patent_Abstract	Similarity_roberta-base-nli-mean-tokens	Similarity_bert-large-nli-mean-tokens	Similarity_allenai_scibert_scivocab_uncased
0	C12P7/14	the efficient production of ethanol from low-c...	0.728632	0.657154	0.691356
1	C12P5/00	the efficient production of ethanol from low-c...	0.728632	0.657154	0.691356
2	C12P5/023	the efficient production of ethanol from low-c...	0.728632	0.657154	0.691356
3	C12P7/06	the efficient production of ethanol from low-c...	0.728632	0.657154	0.691356
4	C12P7/10	the efficient production of ethanol from low-c...	0.728632	0.657154	0.691356
...	...	...	...	...	...
1047	B29C31/006	on a foam plastics molding machine a molding u...	0.873653	0.708035	0.800197
1048	B29C33/303	on a foam plastics molding machine a molding u...	0.873653	0.708035	0.800197
1049	B29C44/428	on a foam plastics molding machine a molding u...	0.873653	0.708035	0.800197
1050	B29C45/1742	on a foam plastics molding machine a molding u...	0.873653	0.708035	0.800197
1051	B29C45/2675	on a foam plastics molding machine a molding u...	0.873653	0.708035	0.800197

Fig. 3. Similarities between patent classes and patent abstracts

	Model	thresholdl_60	threshold_65	threshold_70	threshold_75	threshold_80
0	roberta-base-nli-mean-tokens	100.00%	100.00%	100.00%	97.92%	85.42%
1	bert-large-nli-mean-tokens	100.00%	98.96%	85.42%	55.21%	15.62%
2	allenai/scibert_scivocab_uncased	100.00%	98.96%	91.67%	66.67%	14.58%

Fig. 4. Different Bert models results with different thresholds

• Similarity Model Generalization

The RoBERTa model is used to evaluate similarities between tweet pairs and patent classes, as well as academic summaries and patent classes. We classify, according to the predefined reference threshold, technologies that are already

patented and that are currently of public interest, on social networks or in academic research, and those that can be regarded as new opportunities (those that do not exceed the similarity threshing).

For illustration, as shown in Fig. 5, the content of tweet 8 addresses a technology already patented. It is comparable to 81.19% with class E04C2/205. Both texts are devoted to the use of plastic sheets in building construction, including the installation of polyethylene sheets under concrete tiles to ensure their stability.

Tweet ID: Tweet_8

Class: E04C2/205

Similarity Score: 0.8119176626205444

Tweet Content: RELIABILITY GUARANTEE VE KNOW CHOOSE GOOD INNOVATION RELIABILITY GUARANTEE PLASTIC MANUFACTURER INDUSTRY LEAD INNOVATION RELIABILITY RELIABILITY GUARANTEE MANUFACTURER UK LEAD INDEPENDENT FILM INDUSTRY LEAD INDEPENDENT PLASTIC MANUFACTURER UK LEAD INDEPENDENT VE SUPPLY PACKAGING INDUSTRY FILM PLASTIC MANUFACTURER MANUFACTURER UK KNOW PACKAGING INDUSTRY LEAD CHOOSE GOOD HAND KNOW CHOOSE FILM PACKAGING INDUSTRY POLYTHENE FILM PACKAGING QUALITY INNOVATION RELIABILITY POLYTHENE FILM SUPPLY POLYTHENE GOOD HAND VE SUPPLY POLYTHENE INDEPENDENT FILM PLASTIC FILM PACKAGING UK KNOW CHOOSE FILM PLASTIC INDEPENDENT FILM SUPPLY POLYTHENE FILM GUARANTEE VE QUALITY INNOVATION GUARANTEE VE SUPPLY UK KNOW CHOOSE GOOD

Class Content: BUILDING ELEMENTS OF RELATIVELY THIN FORM FOR THE CONSTRUCTION OF PARTS OF BUILDINGS EG SHEET MATERIALS SLABS OR PANELS -CHARACTERISED BY SPECIFIED MATERIALS -OF WOOD FIBRES CHIPS VEGETABLE STEMS OR THE LIKE; OF PLASTICS; OF FOAMED PRODUCTS -OF PLASTICS-OF FOAMED PLASTICS OR OF PLASTICS AND FOAMED PLASTICS OPTIONALLY REINFORCED

Fig. 5. Example of correspondence (tweet 8 and class E04C2/205)

4.2. LDA Subject Modeling

I explored new technological possibilities by implementing LDA topic modeling on documents identified as promising areas using our RoBERTa model. This approach enabled the identification of emerging topics in the field of plastics research, such as reducing persistent plastic waste and exploring innovative biomaterial substitutes [26], among other advanced materials. This initiative aims to pave the way for new technological advancements in the plastics sector.

To assess the quality and coherence of the topics generated by our LDA model, we calculated the consistency score using the Cv Consistency Method. Optimizing alpha (0.01) and beta (0.9) hyperparameters resulted in a coherence score of 0.527837. These hyperparameters play a crucial role in shaping how topics are distributed across documents and the prominence of words within topics. Fig. 6 illustrates how refining these parameters led to the identification of distinct and coherent topics, offering valuable insight and actionable perspectives.

Thematic coherence checks are essential to ensure that identified topics are not only relevant but also clearly defined, thereby enhancing the reliability of our findings and pinpointing potential technological opportunities. Fig. 7 shows the top 20 words and their respective weights in each topic, providing a detailed view of the thematic composition of the identified topics. Visual presentations facilitate a deeper understanding of dominant themes and their relative significance within topics, aiding in result interpretation and strategic decision-making for exploring new technological horizons.

	Validation_Set	Topics	Alpha	Beta	Coherence
0	75% Corpus	2	0.01	0.01	0.520035
1	75% Corpus	2	0.01	0.31	0.513115
2	75% Corpus	2	0.01	0.61	0.528406
3	75% Corpus	2	0.01	0.9099999999999999	0.527837
4	75% Corpus	2	0.01	symmetric	0.507593
...	...	...	...	...	...
115	100% Corpus	3	asymmetric	0.01	0.487882
116	100% Corpus	3	asymmetric	0.31	0.430966
117	100% Corpus	3	asymmetric	0.61	0.443902
118	100% Corpus	3	asymmetric	0.9099999999999999	0.508055
119	100% Corpus	3	asymmetric	symmetric	0.430966

Fig. 6. Best hyperparameters LDA

```
[(0,
  '0.007*"industry deliver" + 0.007*"great customer experience" + 0.007*"quality product innovation" + 0.007*"solution plastic" + 0.007*"solution plasti
c industry" + 0.007*"customer experience solve" + 0.007*"technical challenge commitment" + 0.007*"provide innovative solution" + 0.007*"challenge commit
ment" + 0.007*"precision craft"'),
 (1,
  '0.008*"exq ceugsi mxjtjs" + 0.008*"exq ceugsi" + 0.008*"daily like" + 0.008*"daily like rabbit" + 0.008*"bullet chinese military" + 0.008*"boast wtf"
+ 0.008*"ceugsi mxjtjs" + 0.008*"innovation come" + 0.008*"hat despite" + 0.008*"come daily"'),
 (2,
  '0.007*"recycle plastic" + 0.007*"sociolla statement" + 0.007*"plastic waste" + 0.007*"plastic waste beauty" + 0.007*"product sociolla statement" + 0.
007*"recycle plastic waste" + 0.007*"recycle product" + 0.007*"new innovation recycle" + 0.007*"understand cost" + 0.007*"understand cost recycle"')]
```

Fig. 7. Topic with word weight

To make the results of the topic modeling more visible and interpretable, we used the interactive tool pyLDAvis. This tool allows for an intuitive exploration of the topics generated by the LDA model. The visualization of pyLDAvis of the two main identified topics is presented in Figure 8. In this visualization, each circle represents a topic, and the distance between the circles indicates the difference between the topics. The further apart the circles are, the more distinct the topics are. Circles of different sizes indicate the importance of the topic within the corpus.

The analysis of this illustration highlights two distinct groups of topics. The main objective of the first group is to focus on the transition to a plastic circular economy in order to change the trend of collision trajectory between plastics and the environment. The second group deals with technologies aimed at reducing the amount of plastic waste by using alternative plastics based on innovative biomaterials. Through this segmentation, it is possible to identify specific areas of research and innovation that offer strategic perspectives for the development of new technological opportunities in the plastics sector.

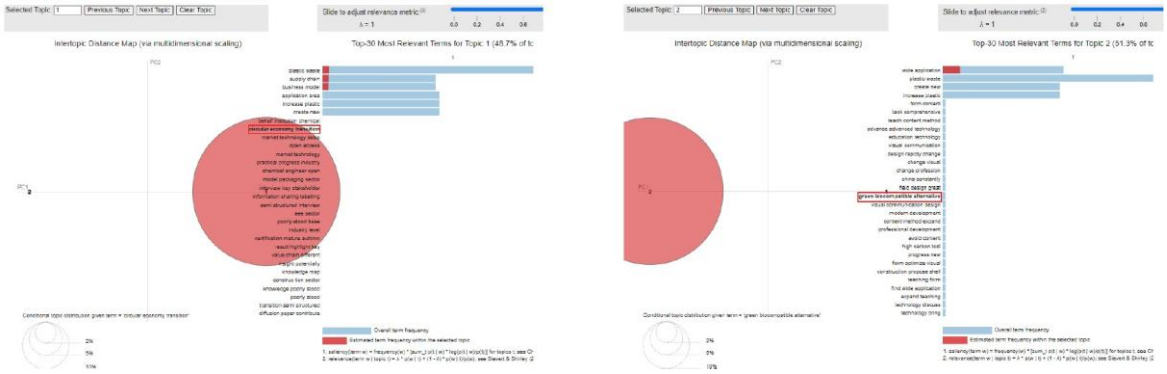

Fig. 8. Visualization of the top-30 most relevant terms for the two main topics identified

The distribution of technology topics in each document is shown in Fig. 9. The emerging technologies and dominant research areas are shown in this graph. It identifies the most promising research topics and potential gaps in the plastics sector by analyzing the distribution of different elements of technological progress across the documents.

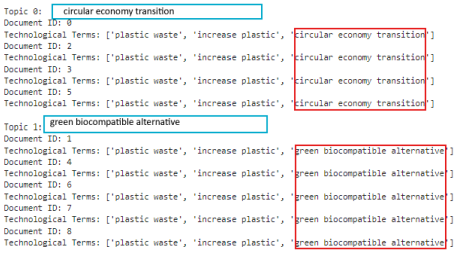

Fig. 9. documents with technological terms per topic

5. CONCLUSION

In summary, this research has demonstrated the effectiveness of using advanced machine learning and natural language processing methods to monitor technological advances in plastics innovation. Using tools such as RoBERTa and LDA, as described in the methodology 3, we have successfully identified new technological trends and potential opportunities in the plastics sector. The results we obtained have highlighted key areas such as reducing plastic consumption, exploring bio-based alternatives, and changing circular economy practices. In addition, cosine similarity analysis and LDA thematic modeling have helped to classify technological concepts, providing strategic guidance for future research and

innovation. Through these methodologies, technological research can be improved, paving the way for sustainable development and innovation in the field of plastics. As technology is constantly evolving, the integration of these cutting-edge approaches ensures a better understanding and exploitation of technological landscapes, enabling proactive and informed decision-making in industrial and environmental contexts.

References

[1] Jérôme Petigny; Christophe Ménigault *et al.* (2019). Étude économique sur l'industrie, les marchés et les déchets du plastique au Canada. Environnement et Changement climatique. https://www.publications.gc.ca/site/fra/9.871297/publication.html.

[2] A Goncalves; G Cardeal *et al.* (2024). Sustainable Value Roadmap for the Plastics Industry. Volume 122, 2024, pp 419-424.

[3] A.L. Porter, S.W. Cunningham (2004), Tech Mining: Exploiting New Technologies for Competitive Advantage.

[4] A. Avila-Robinson, N. Islam, S. Sengoku (2022). Exploring the knowledge base of innovation research: towards an emerging innovation model. Technological Forecasting and Social Change, 182, 121804.

[5] D. Chiavetta, A. Porter (2013). Tech Mining for Innovation Management.

[6] Kobayashi *et al.* (2018). Text classification for organizational researchers: a tutorial. Organizational Research Methods.

[7] Jinfeng Wang *et al.* (2023). Development of technology opportunity analysis based on technology landscape by extending technology elements with BERT and TRIZ. Technological Forecasting and Social Change, 191, 122481.

[8] Choi *et al.* (2013). An SAO-based text-mining approach for technology roadmapping using patent information. R and D Management, 43(1), pp. 52–74.

[9] A. Abbas, L. Zhang, S.U. Khan (2014). A literature review on the state-of-the-art in patent analysis.

[10] S.H. Liu, H.L. Liao, S.M. Pi, J.W. Hu (2011). Development of a patent retrieval and analysis platform–a hybrid approach. Expert Systems with Applications, 38(6), pp. 7864–7868.

[11] Yuming Liu *et al.* (2024). Technology status tracing and trends in construction robotics: A patent analysis. World Patent Information, 76, 102259.

[12] Insu Cho, Yonghan Ju (2023). Text mining method to identify artificial intelligence technologies for the semiconductor industry in Korea. World Patent Information, 74, 102212.

[13] Xinyue Hu *et al.* (2024). Mapping the field: A bibliometric literature review on technology mining. Journal of Pipeline Systems Engineering and Practice, 15(3), 04024017.

[14] Zone-Ching Lin *et al.* (2016). Combination of improved cosine similarity and patent attribution probability method to judge the attribution of related patents of hydrolysis substrate fabrication process. Advanced Engineering Informatics, 30(1), pp. 26–38

[15] Ashkan Ebadi (2022). Detecting emerging technologies and their evolution using deep learning and weak signal analysis. Journal of Informetrics, 16(4), 101344.

[16] Qianwen Ariel Xu *et al.* (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges.

[17] Anil Sharma et Suresh Kumar, (2023). Ontology-based semantic retrieval of documents using Word2vec model.

[18] C. Lee, D. Jeon, J.M. Ahn, O. Kwon (2020). Navigating a product landscape for technology opportunity analysis: a word2vec approach using an integrated patent-product database. Technovation, 96-97, 102140.

[19] Zhang Yi *et al.* (2017). Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016. Knowledge-Based Systems, 133, pp. 255–268.

[20] Zirn et Stuckenschmidt (2014). Analyzing Positions and Topics in Political Discussions of the German Bundestag.

[21] Chen Tian *et al.* (2022). Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model.

[22] Wang, B., S. B. Liu, K. Ding, Z. Y. Liu, and J. Xu. (2014). Identifying Technological Topics and Institution-Topic Distribution Probability for Patent Competitive Intelligence Analysis: A Case Study in LTE Technology. Scientometrics, 101(1), pp. 685–704.

[23] C. Bonten; M. Smith (2019). Plastics Technology: Introduction and Fundamentals.

[24] M.T. MacLean-Blevins (2017). Designing Successful Products with Plastics: Fundamentals of Plastic Part Design, Plastics Design Library.

[25] Mariusz Salwin; Andrzej Kraslawski (2024). Product-Service System business model for plastics industry. Journal of Cleaner Production, 451, 141874.

[26] Nikita Verma *et al.* (2023). Green sustainable biocomposites: Substitute to plastics with innovative fungal mycelium-based biomaterial. Journal of Environmental Chemical Engineering, 11(5), 110396.

CHAPITRE 4 : CONCLUSION GÉNÉRALE

L'objectif de cette conclusion est de résumer les principaux résultats obtenus, de discuter de leur importance dans le contexte plus large de l'industrie des plastiques et de proposer des perspectives pour des recherches futures. Le travail présenté dans ce mémoire se divise en trois chapitres : une introduction générale, un article scientifique détaillant notre méthodologie et cette conclusion générale.

4.1. SYNTHÈSE DES RÉSULTATS

Les fondements théoriques et méthodologiques de notre étude sont posés dans le premier chapitre. Il a été souligné l'importance de la prévision technologique dans le secteur des plastiques, un secteur en constante évolution marqué par des progrès continus et des marchés complexes. Le maintien de la compétitivité et la réponse aux exigences environnementales croissantes nécessitent une anticipation des avancées technologiques pour les entreprises. Ce chapitre met également en évidence l'importance d'une méthode intégrée et automatisée pour analyser les grandes quantités de données disponibles. Nous avons également souligné l'importance de l’apprentissage automatique dans le domaine de la surveillance technologique, qui permet de traiter de manière efficace des ensembles de données complexes et considérables, créant ainsi de nouvelles possibilités pour la découverte d'innovations radicales.

Le chapitre suivant, intitulé État de l'art, examine les études et les connaissances actuelles dans le domaine de la surveillance technologique. Les méthodes de veille technologique actuelles, les approches pour repérer les tendances et les innovations ainsi que les lacunes dans la littérature que cette étude cherche à combler sont abordées dans ce chapitre. Nous examinons les diverses méthodes employées afin de surveiller et d'analyser les données technologiques, en mettant en évidence les avantages et les inconvénients de

chaque méthode. L'analyse de l'état de l'art souligne l'importance d'améliorer les techniques existantes afin de mieux gérer les grandes quantités de données et d'identifier plus précisément les dernières tendances.

Le troisième chapitre de cette recherche se focalise sur une méthode de prévision technologique automatisée, combinant diverses sources de données comme les bases de brevets, les réseaux sociaux et les résumés scientifiques pour offrir une vue d'ensemble des évolutions technologiques. L'analyse sémantique est enrichie à l'aide de techniques avancées de traitement du langage naturel (NLP), telles que l'extraction de Ngrams et des modèles pré-entraînés comme RoBERTa. Divers modèles de similarité sémantique (distilbert-base-uncased, bert-base-nli-mean-tokens) ont été validés, principalement par la similarité cosinus.

L'extraction des informations s'appuie sur des bases comme l'USPTO, collectant de manière structurée des données clés (titres, résumés, revendications, dates de dépôt, inventeurs) grâce à des API. Concernant les médias sociaux, l'API Twitter (via Ntscrapper et Nitter) a permis de récolter et filtrer des tweets pertinents. Les données collectées incluent les textes, mentions, dates et interactions, comme les « j'aime » et les commentaires, telles que récapitulées dans la Table 2, qui montre un aperçu détaillé de ces bases de données et des informations extraites via les API.

Table 2. Bases de données.

Base de données	**API utilisée**	**Nombre de textes extraits**	**Champs de données collectés**	**Période de collecte**
USPTO	PatentsViews	1244	Résumé, CPC_subgroup, CPC title	Aléatoire
Web of Science	Web of Science	31	Titre, résumé	Aléatoire
Twitter (X)	Nitter	48	Lien, texte, date, likes, commentaires	Juin 2024

- Extraction de Ngrams

L'extraction des Ngrams a été réalisée avec la bibliothèque Scikit-Learn via CountVectorizer, permettant de générer des Ngrams de différentes tailles (mono-grams, bi-grams, tri-grams) tout en filtrant les mots vides. Cette approche flexible ajuste les paramètres en fonction des besoins spécifiques de l'analyse. Le processus vise à capturer les concepts technologiques clés ainsi que les associations sémantiques pertinentes. Les résultats de cette extraction, illustrant les Ngrams les plus pertinents, sont résumés dans la Table 3, qui met en lumière les principales associations sémantiques identifiées.

Table 3. Nombre total des Ngrams.

Source	**Nombre des bi-grams extraits**	**Nombre des tri-grams extraits**	**Méthode d'extraction de Ngrams**
Tweets	1084	1057	CountVectorizer
Résumé de brevets	49971	54756	CountVectorizer
Classe de brevet	18000	17809	CountVectorizer
Résumé d’articles	3830	3973	CountVectorizer

Notre méthodologie a été appliquée à quatre corpus distincts : les classes de brevets, les résumés de brevets, les résumés d’articles scientifiques et les tweets. Les résultats obtenus ont montré notre capacité à identifier des motifs textuels pertinents tels que « *cure plastic* », « *welding plastic* », « *plastic recycling* », « *plastic waste* » et « *fiber reinforce plastic* ». Pour des détails supplémentaires, voir la Figure 4.

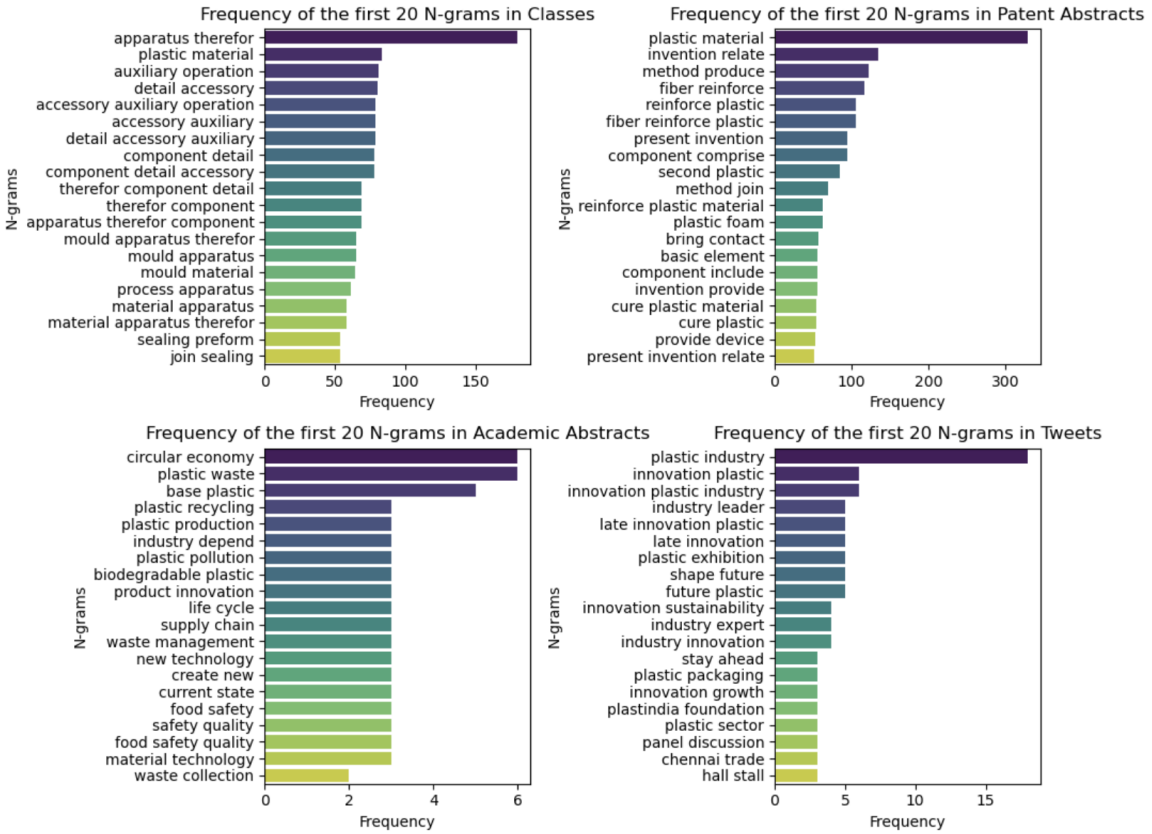

Figure 4. Fréquences des 20 premiers Ngrams.

- **Modèles pré-entraînés**

Nous avons utilisé plusieurs modèles pré-entraînés, tels que RoBERTa, distilbert-base-uncased et bert-base-nli-mean-tokens, pour améliorer la représentation sémantique des textes. Ces modèles ont été ajustés sur notre corpus de données pour optimiser leur performance sur des tâches spécifiques de similarité sémantique. À ces méthodes s'ajoute l'allocation latente de Dirichlet (LDA) pour détecter et extraire des éléments dissimulés dans les ensembles de données textuelles. Les collections de documents ont été divisées en un ensemble de sujets, chacun représenté par un ensemble de mots grâce à la technologie LDA. Cette méthode améliore notre capacité à détecter les tendances émergentes en représentant des sujets cohérents et pertinents dans les données de brevets, les tweets et les résumés d'articles scientifiques. Les conclusions de notre étude sont extrêmement encourageantes.

L'approche suggérée a permis de détecter non seulement les technologies en cours, mais aussi les nouvelles avancées technologiques et les possibilités futures. À titre d'exemple, nous avons pu détecter des évolutions vers des solutions plus durables et respectueuses de l'environnement dans le domaine du plastique, comme les plastiques biodégradables ou recyclés. En identifiant ces sujets à temps, les entreprises bénéficient d'informations stratégiques cruciales, ce qui leur permet de prendre des décisions d'investissement et de développement de produits plus éclairées et proactives.

En analysant les tendances et les sujets en plein essor, notre étude a permis d'établir un aperçu des dernières avancées technologiques. Grâce à l'analyse des similitudes et à l'extraction de thèmes, nous avons pu évaluer comment les technologies émergentes sont perçues et acceptées par la communauté scientifique et le secteur public. Cette approche a mis en lumière non seulement les changements dans les discussions sur de nouvelles matières et technologies, mais aussi également éventuelles résistances ou d'obstacles à leur adoption.

Parmi les résultats marquants de notre analyse, nous avons identifié un intérêt croissant pour des solutions écologiques et biocompatibles, notamment l'utilisation du mycélium fongique comme alternative au plastique. Issu des champignons, le mycélium fongique constitue une option biodégradable et durable pour remplacer de nombreux matériaux plastiques. Selon Nikita Verma *et al*. (2023), il offre des caractéristiques remarquables : il est léger, solide et peut être modelé sous différentes formes, ce qui le rend adapté à une large gamme d'applications, allant des emballages aux matériaux de construction. Ce sujet, détecté grâce à des méthodes avancées de clustering et de détection de sujets, présente un potentiel significatif pour transformer l'industrie du plastique. Bien que ce thème n'ait pas été abordé en profondeur dans notre article scientifique en raison de contraintes d'espace, il représente une opportunité majeure d'innovation qui mérite une attention particulière. En complément, des approches issues de l'économie circulaire ont également été identifiées pour améliorer la gestion des déchets plastiques, notamment la technologie Pyrowave. Cette innovation repose sur la dépolymérisation par micro-ondes pour recycler les plastiques usagés en monomères, permettant ainsi une réutilisation infinie sans perte de qualité. Bien que cette technologie n'ait pas été également intégrée dans notre article en raison de limitations d'espace, elle

constitue une avancée majeure pour promouvoir une gestion durable et efficace des déchets plastiques.

En résumé, notre approche méthodologique intégrée offre une meilleure compréhension des dynamiques technologiques dans le domaine des plastiques et permet de repérer les innovations potentielles avant qu'elles ne soient brevetées. Il est essentiel que les entreprises aient cette capacité d'anticipation afin de maintenir leur compétitivité et de diriger leurs investissements de manière stratégique. La recherche renforce ainsi la surveillance technologique et offre une base solide pour l'innovation stratégique dans un domaine en perpétuelle mutation.

4.2. IMPORTANCE ET IMPLICATION DES RÉSULTATS

Les résultats obtenus ont de nombreuses conséquences significatives pour le secteur des plastiques. Tout d'abord, l'utilisation d'une méthode intégrée et automatisée permet de gérer de manière efficace de grandes quantités de données tout en offrant des informations plus précises et exploitables sur les évolutions technologiques. Cela revêt une importance capitale pour les entreprises qui souhaitent maintenir leur compétitivité et innover en réponse aux changements du marché. Ensuite, l'utilisation des modèles de similarité sémantique dans cette étude, en collaboration avec l'allocation latente de Dirichlet (LDA), a prouvé leur aptitude à repérer des liens subtils entre les concepts technologiques et à extraire des sujets spécifiques. Il est crucial d'avoir cette capacité afin de prévoir les orientations futures de l'innovation et de repérer les technologies susceptibles de dominer.

En ce qui concerne la validité et la fiabilité de nos résultats, nous avons une méthodologie rigoureuse et des techniques de validation standardisées. Cependant, il convient de prendre en compte les contraintes de notre étude. À titre d'exemple, l'utilisation des sources de données disponibles et la complexité des algorithmes de traitement du langage naturel peuvent engendrer des préjugés, ce que nous n’avons pas pu évaluer. De plus, il est

possible que certaines informations ne soient pas à jour ou ne soient pas pertinentes, voire fausses, ce qui pourrait avoir un impact sur les résultats de l'analyse.

4.2.1. Perspectives de recherches futures

Les opportunités de recherche à venir sont multiples et prometteuses. Dans un premier temps, la progression constante des modèles de traitement du langage naturel, en particulier avec les modèles les plus récents et les plus performants, pourrait encore améliorer la précision et la pertinence de nos analyses. Les performances pourraient être encore améliorées grâce à des modèles tels que GPT-4 ou les futures itérations de BERT.

Il est également intéressant d'élargir le champ des sources de données. Il serait possible d'intégrer des sources comme les forums spécialisés, les salons et les conférences industrielles afin d'obtenir une vision encore plus exhaustive des évolutions technologiques. Cela offrirait une vision plus étendue et plus variée, comprenant des points de vue et des tendances qui ne pourraient pas être prises en compte par les sources de données classiques.

De plus, l'incorporation d'autres méthodes d'intelligence artificielle, telles que l'apprentissage profond et les réseaux de neurones, pourrait offrir de nouvelles possibilités pour l'analyse prédictive et la détection des nouvelles avancées technologiques. Prenons par exemple l'emploi de réseaux de neurones convolutifs (CNN) pour analyser des images de brevets ou des graphes de citation, cela pourrait apporter des informations supplémentaires. Par ailleurs, une étroite collaboration avec des spécialistes de l'industrie des plastiques pourrait contribuer à donner un contexte aux résultats et à orienter les recherches à venir vers des applications concrètes et directement exploitables par les acteurs industriels. Grâce à des collaborations avec des entreprises et des institutions de recherche, il serait possible de confirmer les résultats dans des situations concrètes et de concevoir des outils et des solutions directement appliqués.

4.3. CONCLUSION FINALE

En résumé, cette étude a mis en évidence l'efficacité d'une méthode dynamique, intégrée et automatisée pour prédire les avancées technologiques dans le secteur industriel des plastiques. Les résultats obtenus valident les hypothèses initiales et démontrent que notre approche peut offrir des connaissances précieuses pour orienter les choix stratégiques et les investissements dans cette industrie en constante évolution.

Les méthodes et les modèles élaborés dans cette étude offrent les fondations pour des recherches ultérieures sur les technologies émergentes. En analysant les données provenant des réseaux sociaux et des publications scientifiques, ces recherches permettront de repérer les avancées technologiques non encore brevetées, tout en favorisant le développement de systèmes de veille proactifs pour anticiper les tendances dans divers secteurs stratégiques. De plus, l'expansion des sources de données, comme l'intégration de plateformes professionnelles telles que LinkedIn ou de bases de données techniques comme le Material Data Center, permettrait de capter des signaux encore plus précis. L'amélioration des modèles actuels, en intégrant par exemple des architectures basées sur GPT-4 pour une meilleure compréhension des tendances sémantiques, faciliterait le développement de systèmes proactifs capables de proposer des recommandations stratégiques aux entreprises, en particulier dans le cadre de la transition vers une économie circulaire.

ANNEXE : DEUXIÈME ARTICLE

Exploitation des grands modèles de langage pour l'extraction précise de sujets et la nomination de brevets technologiques : une méthodologie centrée sur GPT.

Résumé en français du deuxième article

La modélisation thématique est largement utilisée pour analyser de vastes ensembles de données textuelles, notamment dans le cadre de la nomination de brevets technologiques. Cependant, les méthodes traditionnelles, telles que la Latent Dirichlet Allocation (LDA) et la Factorisation en Matrice Non-Négative (NMF), présentent des limites en termes de compréhension sémantique, de scalabilité et d'adaptabilité. Cette étude explore l'utilisation des Grands Modèles de Langage (LLMs) pour surmonter ces défis. En exploitant leur compréhension avancée du langage, les modèles GPT génèrent des thématiques cohérentes et contextuellement pertinentes, surpassant les méthodes conventionnelles. Ils se distinguent également par leur capacité à identifier les tendances émergentes, améliorant ainsi les processus de nomination de brevets. Nos résultats mettent en lumière le potentiel des approches basées sur GPT pour simplifier l'analyse des brevets et accélérer la découverte technologique.

Mots-clés : Innovation technologique, modélisation thématique, TopicGPT, GPT, LDA.

Cet article, intitulé « *Harnessing Large Language Models for Precision Topic Extraction and Technology Patent Nomination : A GPT-centric Methodology*», a été soumis pour publication à la conférence EDI40-25 2024, *The 8th International Conference on Emerging Data and Industry*, qui se tiendra du 22 au 24 avril 2025 à Patras, Grèce (Tshibanda *et al.*, 2024). En tant que premier auteur, j'ai contribué principalement à la recherche sur l'état de la question, au développement de la méthodologie et à son opérationnalisation. Said

Echchakoui, second auteur, a aidé à la recherche sur l'état de la question, au développement de la méthode ainsi qu'à la révision de l'article. Adda Mehdi, le troisième auteur, a fourni l'idée originale, a aidé à la recherche sur l'état de la question et a également contribué à la revue de la littérature.

Harnessing Large Language Models for Precision Topic Extraction and Technology Patent Nomination: A GPT-centric Methodology

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Data and Industry (EDI40), April 22-24, 2025, Patras, Greece

Harnessing Large Language Models for Precision Topic Extraction and Technology Patent Nomination: A GPT-centric Methodology

Franck Tshibanda Nkolongo[a,*], Said Echchakoui[b], Adda Mehdi[a]

[a]*Department of Mathematics, Computer Science and Engineering, University of Quebec at Rimouski, Canada*
[b]*Management Sciences Departmental Unit, University of Quebec at Rimouski (Lévis), Canada*

Abstract

Topic modeling is widely used for analyzing large textual datasets, particularly in technology patent nomination. Traditional methods, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), often struggle with semantic comprehension, scalability, and adaptability. This study investigates Large Language Models (LLMs) to overcome these limitations. Leveraging their advanced language understanding, GPT models generate coherent and contextually relevant topics, outperforming conventional methods. They also excel in identifying emerging trends, enhancing patent nomination processes. Our findings demonstrate the potential of GPT-based approaches to streamline patent analysis and accelerate technological discovery.

Keywords: Patent Analysis; Topic Modeling; Large Language Models (LLMs); GPT Applications; Natural Language Processing (NLP)

1. Introduction

Topic modeling is a crucial method for analyzing large textual datasets, especially in technology patent nomination, as it reveals hidden thematic structures that help identify emerging trends and classify innovations. However, traditional techniques such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) face significant limitations, particularly in handling the complexities of patent language, including polysemy and synonymy [3][4]. These challenges often result in less coherent and contextually relevant topics.

LDA assumes word independence, which undermines its effectiveness in specialized domains like patents [2]. Although NMF is interpretable, it requires extensive parameter tuning and is sensitive to initialization, making it unreliable for dynamic datasets [9]. Both models are also static, necessitating complete reruns when new documents are added, which is inefficient for continuously evolving patent repositories [3][6].

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: author@institute.xxx

Given these limitations, there is a growing need for more sophisticated models. Large Language Models (LLMs), particularly Generative Pre-trained Transformers (GPT), have emerged as promising alternatives. LLMs, such as GPT-3 and GPT-4, can capture semantic relationships and do not rely on the bag-of-words assumption, making them well-suited for technical language [12][11]. Their ability to generate contextually relevant topics without extensive retraining offers significant advantages over traditional methods [16].

This study aims to explore how LLMs, specifically GPT, can enhance topic extraction in patent analysis. It focuses on two objectives: 1) evaluating how GPT models improve the accuracy and coherence of extracted topics compared to traditional techniques, and 2) assessing the effectiveness of GPT-driven modeling in identifying and nominating new technology patents.

The main research questions explored in this study are: How do LLMs, particularly GPT models, enhance the accuracy and coherence of topic extraction in patent analysis? Additionally, can GPT-driven topic modeling effectively identify and nominate new technology patents, and how do these models compare to traditional methods The study compares the effectiveness of GPT models for topic modeling in patent analysis against traditional methods like LDA and NMF. The approach leverages the capabilities of LLMs to extract relevant topics from patent abstracts, taking into account technical linguistic specifications. The results show that GPT models outperform traditional methods in terms of accuracy and coherence, allowing for better identification of emerging technologies and more precise titles for new patents.

The remainder of this paper is structured as follows. Section 2 provides a review of traditional techniques and LLMs. Section 3 outlines the methodology and data collection process. Section 4 presents the results, and Section 5 discusses the findings, focusing on the identification of new patent classes and refining LLM-based topic extraction models.

2. Literature Review

Topic modeling has long been a crucial tool for discovering latent topics in text corpora, particularly in fields such as scientific papers and patents. Among the traditional methods most widely used, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) stand out.

LDA, introduced by [3], models documents as mixtures of topics. However, LDA relies on the ”bag-of-words” assumption, which does not account for dependencies between terms within a document. This limitation makes LDA less effective for specialized corpora, such as patents, where technical language and context are critical [2]. Moreover, LDA requires prior specification of the number of topics, a computationally expensive and rigid process, especially with dynamic datasets.

NMF, another widely used method, decomposes the term-document matrix into non-negative factors, providing better interpretability than LDA. However, like LDA, NMF struggles to capture semantic nuances and lacks adaptability when dealing with evolving datasets [2, 3].

In a study by [17], LDA was combined with models like BERT to assess semantic similarity in patents. While embedding-based approaches provided better textual representations than LDA alone, several challenges remained. One of the main issues was the difficulty in capturing granular details of technical concepts, and LDA failed to effectively address the interconnectedness of these concepts. Additionally, naming the identified topics with precision proved problematic, underscoring the limitations of traditional methods when applied to highly specialized fields.

The emergence of Large Language Models (LLMs), such as Generative Pre-trained Transformers (GPT), has led to a paradigm shift in text analysis. Based on the [10] architecture, LLMs overcome many of the limitations associated with traditional methods by capturing complex semantic relationships and producing more coherent topic structures. Trained on vast corpora of text, models like GPT-3 and GPT-4 excel in understanding specialized languages, making them particularly suitable for analyzing patent data [11].

In contrast to traditional methods, LLMs offer remarkable flexibility through their zero-shot and few-shot learning capabilities, allowing them to rapidly adapt to new data without the need for full retraining [12]. Prompt-based methods further facilitate precise topic extraction [11]. The TopicGPT framework, developed by [14], exemplifies this flexibility, enabling iterative semantic refinements, in contrast to systems like GoalEx [7], which focus solely on corpus partitioning.

Another study by [18] applied few-shot learning in LLMs to multilingual patent data, highlighting the models' ability to uncover latent translingual topics. This demonstrates the versatility of LLMs in handling linguistically diverse datasets, further emphasizing their relevance for patent analysis across different languages.

In a comparative analysis by [19], the interpretability of models such as LDA, BERTopic, and RoBERTa was examined, incorporating the capabilities of LLMs into their analysis. They demonstrated that BERTopic and RoBERTa provided more coherent and interpretable topics compared to LDA, largely due to the rich semantic representations offered by LLMs, which are better at capturing the complexity of relationships between topics. The authors also suggested exploring probabilistic models like LSA and PLSA, while highlighting the importance of LLMs for future research in topic modeling, especially in technical domains such as patent analysis.

In conclusion, while traditional methods like LDA, NMF, and BERT have laid a solid foundation for topic modeling, their limitations become evident when applied to specialized corpora like patents. LLMs, on the other hand, offer groundbreaking capabilities for analyzing complex datasets, providing more coherent and interpretable topics while overcoming many of the constraints faced by traditional models. However, optimizing and integrating these models into specific applications remains a promising direction for future research.

3. Methodology

3.1. Data Collection and Preprocessing

This study uses two datasets for comparison. The first dataset includes patents from the United States Patent and Trademark Office (USPTO), spanning 1976–2024, with a focus on the plastics industry due to its relevance to sustainability, recycling, and performance advancements. Initially comprising 65,620 documents, preprocessing reduced this to 9,438 unique patents.

The second dataset consists of 1,434 recent tweets discussing innovations and trends in plastics. Both datasets underwent preprocessing steps, including text cleaning, deduplication, and tokenization, to ensure consistency. These refined datasets were analyzed to compare thematic patterns between patented technologies and emerging discussions on social media.

3.2. Implementation of LLM for Topic Modeling

In recent years, a wide range of LLMs have been developed and deployed, demonstrating remarkable capabilities in NLP tasks. In this study, we focus specifically on GPT-3.5 and GPT-4 for topic extraction due to their advanced language comprehension and ability to handle large, complex datasets. According to [11], these models outperform traditional approaches by generating coherent and contextually relevant topics.

A prompt-based strategy was adopted to guide the models in topic generation and patent naming. Key prompts include:

- Topic Generation Prompt: "Analyze the following patent abstract and generate 5 concise topic names that capture the key technological concepts."
- Topic Refinement Prompt: "Given the following set of topics, suggest more granular sub-topics that could represent potential areas for new patents."
- Patent Naming Prompt: "Based on the identified topics and sub-themes, propose 3 new potential patent domains that address emerging technological needs."

These prompts were applied to evaluate performance in topic extraction and patent naming, emphasizing the role of well-designed instructions in improving accuracy and contextual relevance, as suggested in [11].

To further enhance results, the TopicGPT package was utilized for thematic extraction from preprocessed data. Fine-tuning on domain-specific patents enabled a deeper understanding of technical terms. This implementation eliminated the need to predefine the number of topics, ensuring flexibility and adaptability, as illustrated in Fig. 1 and discussed by [13].

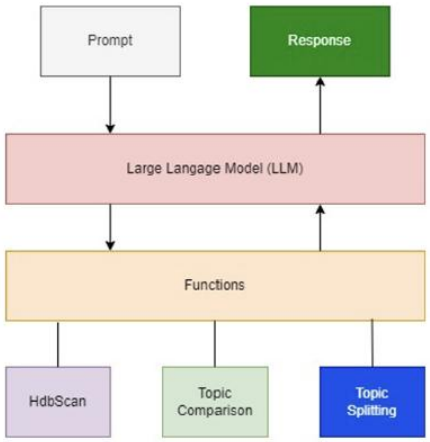

Fig. 1. The chat-based interface for GPTopic (Arik Reuter et al. 2024)

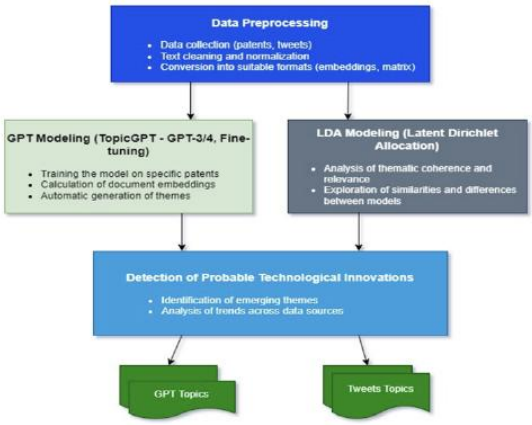

Fig. 2. Methodological Approach

Furthermore, we established a method to compare the themes generated by GPT across various data sources, facilitating the identification of potential technological innovations. Our approach prioritizes the detection of emerging trends and innovations that have yet to be patented (refer to Fig. 2 for a visual representation of this comparative framework).

3.3. Topic Refinement and Patent Nomination

This study employed an automated approach for topic refinement and patent nomination, leveraging GPT models to ensure scalability and reproducibility without relying on human expert review [11].

The refinement process began with generating initial topics from patent abstracts, followed by iterative refinement into more granular subtopics using the Topic Refinement Prompt. The final topics were aggregated into a cohesive model capturing key technological innovations.

We implemented a structured, model-driven nomination process to identify new areas for patent development. This involved:

- Topic-Based Mapping: Analyzing patent distributions to highlight areas with varying concentrations.
- Gap Analysis: Using prompts to uncover underexplored technological domains.
- Novelty Assessment: Evaluating relevance and uniqueness through cosine similarity of document embeddings.

This streamlined approach demonstrates the potential of LLMs to detect emerging trends and propose novel patentable innovations autonomously.

3.4. Evaluation Metrics

We evaluated model performance using five key metrics: **Topic Coherence**, measured by Normalized Pointwise Mutual Information (NPMI) to ensure semantic relevance [2]; **Topic Diversity**, assessed through cosine similarity variance to capture a broad range of topics [15]; **Topic Distinctivity**, which emphasized uniqueness and minimized redundancy [4]; **Patent Nomination Accuracy**, tested against a holdout set of recent patents to validate the identification of novel areas; and **Thematic Similarity**, comparing themes across models to highlight trends and gaps in innovation.

We implemented a traditional LDA model using the same preprocessed dataset, enabling a direct comparison between the GPT-driven approach and traditional methods across all metrics. This comprehensive evaluation underscores the potential of GPT-based models in topic modeling and patent nomination, particularly within the context of technological innovation.

4. Results

4.1. Presentation of Topics and Clusters

The thematic analysis performed using the TopicGPT model, fine-tuned on the GPT-4 architecture, revealed significant insights into both the patent and Twitter datasets. A total of 50 distinct topics were identified within the patent data, illustrating the diverse range of technological innovations present across various sectors (see Fig. 3).

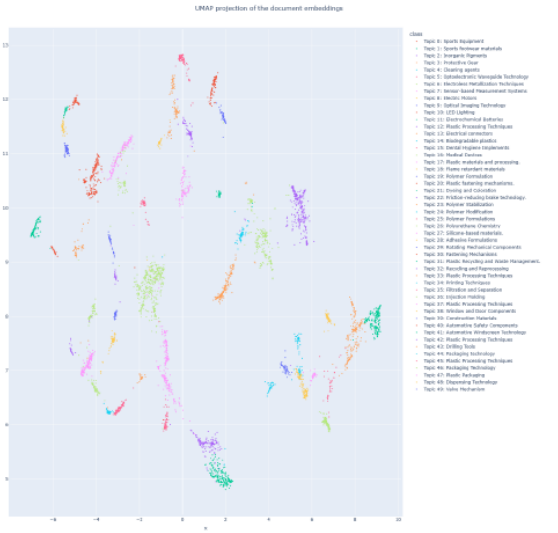

Fig. 3. Clusters of Identified USPTO Topics

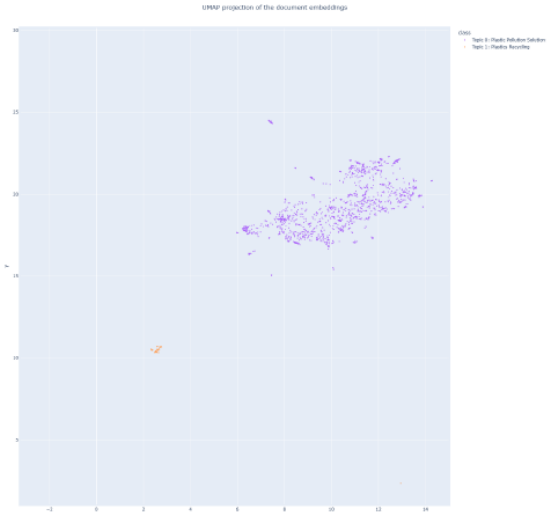

Fig. 4. Clusters of Identified Twitter Topics

Each identified topic corresponds to a unique technological domain, underscoring the complexity and multi-dimensional nature of patent applications. The ability of the TopicGPT model to extract these themes without the need for predefining the number of topics demonstrates its adaptability, making it well-suited for analyzing large and intricate datasets.

In contrast, the analysis of the Twitter data resulted in the identification of 2 major themes. This disparity in the number of topics extracted from the two datasets can be attributed to the varying nature and depth of information present in patents versus tweets. While patents often encompass detailed technical descriptions and claims, tweets tend to be shorter (Fig. 4).

To demonstrate the model's ability to generate relevant thematic areas, we provide examples in Fig. 5 and 6, where one topic is selected from each source, along with its description generated through a prompt.

Topic USPTO (Patents): Sports Footwear Materials This topic encompasses various sub-topics related to sports footwear materials, focusing on aspects such as insoles and soles, sports and outdoor activities, shoe construction and components, traction and grip, and foot protection and comfort. Below is a summary of the key elements associated with this topic:

1) Insoles and Soles: Key terms include insole, toe, resilience, fluff, tread, soles, midsole, arch, and cushioning. 2) Sports and Outdoor Activities: This sub-topic covers terms related to various activities, such as skis, snow, walking, skate, skiing, sports, hockey, and climbing. 3) Shoe Construction and Components: Important components of shoe design include last, shoes, pre-formed uppers, lacing, vamp, heels, and pegs. 4) Traction and Grip: Terms such as spike, studs, traction, cleat, ground-engaging, dig, and grip are critical for enhancing footwear performance. 5) Foot Protection and Comfort: This aspect emphasizes features like padded materials, cushioning, protection, comfort, and shock absorption. Please note that this summary is based on the provided terms and may not encompass the entire range of concepts within the topic.

Topic Twitter (Social Media): Plastic Pollution Reduction This topic addresses the environmental impact of the plastics industry and innovations related to reducing plastic pollution. The various aspects and sub-topics associated with this topic include: 1) Environmental Impact: Key terms encompass pollution, rivers, garbage, recyclability, and toxins. 2) Regulations and Policies: This sub-topic focuses on the role of governments, regulations, negotiations, and associated risks. 3) Climate Crisis: Related concepts include climate action, coal, greenhouse gas emissions, decarbonization, and overall emissions reduction. 4) Health and Safety: Important terms in this area involve risks, exposure, diseases, human health, and hazardous materials. 5) Sustainable Solutions: This aspect highlights approaches such as reduction, recycling, biobased alternatives, the circular economy, and zero waste initiatives.

```
tm_uspto.pprompt("what is the topic 1?")

GPT wants to to call the function:  FunctionCall(arguments='{\n  "topic_idx_lis": [1]\n}', name='get_topic_information')
Topic 1 is about sports footwear materials. Some sub-topics and aspects of this topic include:

1. Insoles and Soles: Insole, toe, resilience, fluff, tread, soles, midsole, arch, cushioning.
2. Sports and Outdoor Activities: Skis, snow, walking, skate, skiing, sports, hockey, climbing.
3. Shoe Construction and Components: Last, shoes, pre-formed, uppers, lacing, vamp, heels, pegs.
4. Traction and Grip: Spike, studs, traction, cleat, ground-engaging, dig, grip.
5. Foot Protection and Comfort: Padded, cushioning, protection, comfort, shock absorption.
```

Fig. 5. Detailed Topic Example from USPTO)

```
tm_twitter.pprompt('What is the topic 1?')

GPT wants to to call the function:  FunctionCall(arguments='{\n  "topic_idx_lis": [1]\n}', name='get_topic_information')
Topic 1 is about Plastic Pollution Reduction. It includes aspects and sub-topics such as Recycling and Waste Management, Environmental Impact, Sustainable Technologies, Innovations in the Plastics Industry, and Government and Regulations. Some of the keywords associated
with this topic are recycleforchange, reduction, dispose, garbage, sorting, polluting, forest, greenhouse, wildlife, marine, solar, renewable, eco-friendly, bio-based, electric, automation, robotics, 3D printing, advanced materials, regulations, policies, regulatory, com
pliance, and standards. [Document 1]
```

Fig. 6. Detailed Topic Example from USPTO)

Additionally, Fig. 7 illustrates the clusters formed from both datasets (tweets and USPTO patents), reduced to two dimensions using the UMAP method. The points in this visualization represent their respective embeddings, enabling observation of data distribution and identification of similarities and differences between the two datasets. Overlapping areas of blue (patents) and green (tweets) points may indicate emerging technologies or innovations discussed in both social media and patent literature, suggesting potential innovative trends and technological development opportunities.

4.2. Evaluation Metrics

The evaluation of the TopicGPT model (GPT-3.5 and GPT-4) reveals several key advantages over traditional LDA in thematic analysis: The results for coherence, diversity, and thematic disctinctiveness are summarized in the following table:

Table 1 presents the results of the thematic analysis conducted using both LDA and TopicGPT, highlighting coherence and thematic distinctiveness scores derived from the USPTO and Twitter datasets. - Coherence: TopicGPT achieves coherence scores of 0.6578 for the USPTO dataset and 0.7313 for the Twitter dataset, significantly outperforming LDA, which shows scores of 0.5227 and 0.5117 (Tshibanda et al., 2024). This indicates that TopicGPT generates more structured and relevant themes, enhancing conceptual alignment within each theme. - Thematic Distinctiveness: Two aspects of distinctiveness were evaluated: -Intertopic Distinctiveness: TopicGPT scores 0.4993 for

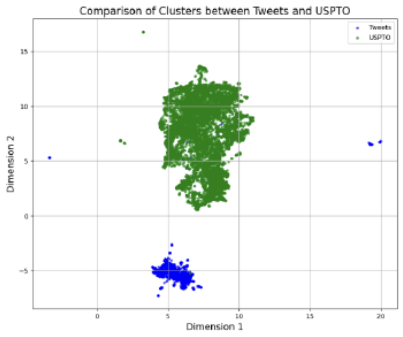

Fig. 7. Comparison of Clusters between Tweets and USPTO Patents

Table 1. Thematic Coherence.

Method	Coherence Uspto	Coherence Twitter	Topic Distinctiveness Twitter	Topic Distinctiveness Uspto	Topic Distinctiveness Twitter vs Uspto	Patent Nomination Accuracy
LDA	0.5227	0.5117	0.2145	0.2029	1.0	0.5267
TopicGPT	0.6578	0.7397	0.4993	0.3119	0.26	0.7823

Twitter and 0.3119 for the USPTO, demonstrating better separation of topics within each dataset compared to LDA's lower scores of 0.2145 (Twitter) and 0.2029 (USPTO). This suggests that TopicGPT captures a broader variety of relevant subjects and reduces overlap between themes. - Distinctiveness Between USPTO and Twitter: LDA achieves a score of 1.0, indicating no overlap between topics from the two datasets, which may oversimplify thematic relationships. In contrast, TopicGPT scores 0.26, suggesting notable overlap between patent themes and social media discussions. This indicates that innovations captured in patents resonate in public discourse, revealing ongoing relevance and interest. - Patent Nomination Accuracy: TopicGPT scores 0.7823 in patent nomination accuracy, significantly higher than LDA's 0.5267. This demonstrates that TopicGPT not only generates coherent and distinct themes but also accurately identifies relevant patents, enhancing its utility in thematic analysis. Fig. 8 presents the Data Frame resulting from the cosine similarity analysis between topics extracted from Twitter and those from USPTO patents. Each row in this table lists a Twitter topic alongside a corresponding USPTO topic, accompanied by a cosine similarity score representing their semantic proximity. The topics are compared based on their vector representations, derived from embeddings calculated for each corpus. The similarity score, ranging from 0 to 1, quantifies this proximity: the closer the score is to 1, the more similar the two topics are thematically.

	twitter_topic	uspto_topic	similarity
0	Topic 0: Plastics industry innovations\n	Topic 0: Sports Equipment\n	0.757564
1	Topic 0: Plastics industry innovations\n	Topic 1: Sports footwear materials\n	0.761885
2	Topic 0: Plastics industry innovations\n	Topic 2: Inorganic Pigments\n	0.724207
3	Topic 0: Plastics industry innovations\n	Topic 3: Protective Gear\n	0.748096
4	Topic 0: Plastics industry innovations\n	Topic 4: Cleaning agents\n	0.739388
...	...	...	...
95	Topic 1: Plastic Pollution Reduction\n	Topic 45: Plastic Processing Techniques\n	0.711582
96	Topic 1: Plastic Pollution Reduction\n	Topic 46: Packaging Technology\n	0.774837
97	Topic 1: Plastic Pollution Reduction\n	Topic 47: Plastic Packaging\n	0.716175
98	Topic 1: Plastic Pollution Reduction\n	Topic 48: Dispensing Technology\n	0.726861
99	Topic 1: Plastic Pollution Reduction\n	Topic 49: Valve Mechanism\n	0.723363

Fig. 8. Similarity between Tweets and USPTO Patents

On average, the observed similarities hover around 75%, suggesting notable correspondence between certain technological discussions on Twitter and innovations already patented in the USPTO database. High-similarity topics indicate a strong overlap, implying that social media discussions of innovations are often already covered by patents. Conversely, lower scores highlight emerging themes or innovation opportunities not yet explored within the legal framework of patents. This differentiation provides valuable insight into areas where technological breakthroughs may emerge from recent social media discussions, while the 75% similarity suggests that there is still room for unpatented innovation.

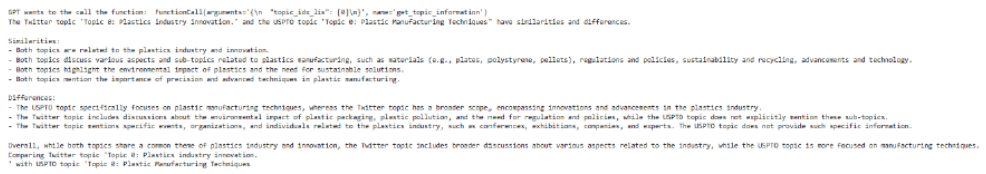

Fig. 9. Example prompting: Topic Similarity between Tweets and USPTO Patents.

As demonstrated in Fig.9, prompts were used to compare themes across different data sources, such as Twitter and USPTO patent abstracts, by identifying semantic similarities and differences. This method allowed for the exploration of connections between social media discussions and patented innovations, revealing emerging technological trends and innovation opportunities. The use of prompts enhanced the speed and flexibility of data exploration, enabling the identification of new areas of interest within the patent landscape. TopicGPT proved to be more effective than LDA, producing more coherent and distinct topics with higher coherence scores for both the USPTO (0.6578) and Twitter (0.7313) datasets. Unlike LDA, which separated topics completely between sources, TopicGPT showed a meaningful overlap (distinctiveness score of 0.26), suggesting that patented innovations influence social media discussions. In terms of Patent Nomination Accuracy, TopicGPT also outperformed LDA (0.7823 vs. 0.6723), reinforcing its value for patent analysis and identifying technological trends.

5. Discussion

This study provides strong evidence for the superiority of GPT-driven approaches in topic modeling and patent nomination, aligning with prior research on the efficacy of LLMs.

One of the key findings is the enhanced semantic understanding demonstrated by GPT models. With coherence scores ranging from 0.65 to 0.73, these models outperformed traditional methods like LDA. Their ability to capture complex relationships in patent documents is particularly valuable in technical domains where language tends to be intricate and specialized. This semantic depth allows for more accurate and contextually relevant topic extraction.

Another notable improvement is in topic diversity and coverage. GPT models exhibited higher distinctiveness scores, ranging from 0.31 to 0.49, compared to LDA. This indicates their superior ability to identify nuanced and emerging topics, reflecting adaptability to dynamic environments. Such adaptability is essential for analyzing evolving technological trends and identifying gaps in innovation strategies.

Additionally, GPT models demonstrated significant advantages in identifying novel patentable areas. With an accuracy rate of 0.78, they far outperformed LDA's 0.52, highlighting their ability to uncover innovation opportunities by synthesizing insights from diverse data sources [16]. This capability underscores their potential as powerful tools for driving patent analysis and fostering innovation.

6. Conclusion and Future Directions

This study demonstrates that large language models are an interesting alternative to traditional methods in topic modeling and patent nomination, offering improvements in semantic coherence, topic diversity, and the identification of new patent areas.

However, challenges remain, including high computational costs that may limit adoption among smaller organizations, as well as concerns about transparency due to their "black box" nature. Future research should address these

limitations by focusing on cost-effective fine-tuning, exploring hybrid approaches to improve efficiency and explainability, and enhancing transparency through techniques such as attention visualization.

References

[1] Yida Mu, *et al.* (2024). "Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling." *LREC-COLING 2024. arXiv:2403.16248.*

[2] Vayansky, I., and Kumar, S.A.P. (2020). A review of topic modeling methods Information Systems 94 (2020) 101582.

[3] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

[4] Grootendorst, M. (2022). BERTopic: Neural Topic Modeling with a Classifier.*arXiv preprint arXiv:2203.00760.*.

[5] Blei, D. M., Lafferty, J. D. (2006). Dynamic Topic Models. Proceedings of the 23rd International Conference on Machine Learning , (pp. 113–120). ACM.

[6] Wang, X., al. (2008). Topical N-grams: Phrase and topic discovery, with an application to information retrieval. In Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), Omaha, Nebraska, USA.

[7] Wang, Z., Levine, S., Reiter, N. (2023). GOALEX: Goal-Driven Explainable Clustering. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

[8] Mu, T., Zhang, L., Li, J. (2024). Large Language Models in Patent Analysis: Overcoming the Limits of Traditional Topic Modeling. Journal of Computational Linguistics.

[9] Lee, D. D., Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

[10] Vaswani, al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

[11] Mu, Y., Li, W., Gao, X. (2024). Advances in Large Language Models for Patent Analysis: A Review and Future Directions. Journal of Artificial Intelligence Research.

[12] Brown, al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[13] Reuter, A., al. (2024). GPTopic: Dynamic and Interactive Topic Representations. Munich Center for Machine Learning (MCML).

[14] Chau Minh, P., Alexander, H. (2024). TopicGPT: A Prompt-based Topic Modeling Framework.

[15] Zhao, W., Zhu, S. (2020). "Topic Diversity in Topic Models: A Measure and Its Applications." *IEEE Transactions on Knowledge and Data Engineering*, 32(4), 821-835.

[16] Zhang, X., Ju, T., Liang, H., Fu, Y., Zhang, Q. (2024). LLMs Instruct LLMs: An Extraction and Editing Method. arXiv preprint arXiv:2403.15736.

[17] Tshibanda, F., Mehdi, A., Echchakoui, S. (2024). Application of machine learning in technological forecasting. Procedia Computer Science Volume 251, 2024, Pages 23-30.

[18] Chen, Y. (2022). Few-shot learning for multilingual patent analysis using large language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9019–9052, Abu Dhabi, United Arab Emirates.

[19] El-Gayar, O., *et al.* (2024). A Comparative Analysis of the Interpretability of LDA and LLM for Topic Modeling: The Case of Healthcare Apps. AMCIS 2024 Proceedings. 22.

RÉFÉRENCES BIBLIOGRAPHIQUES

Alan L. Porter & Michael J. Detampel. (1995). Technology opportunities analysis. Technological Forecasting and Social Change, Volume 49(3), (pp 237–255).

Robert C. Brown *et al*. (2022). Expanding plastics recycling technologies: chemical aspects, technology status and challenges. DOI: 10.1039/D2GC02588D (Tutorial Review) Green Chem, Volume 24, 8899-9002.

Anestis Kousis & Christos Tjortjis. (2023). Investigating the Key Aspects of a Smart City through Topic Modeling and Thematic Analysis. *Future Internet* 2024, Volume*16*(1), 3; https://doi.org/10.3390/fi16010003.

Black, S., & Green, A. (2020). Predicting Technological Trends with USPTO Data. Journal of Intellectual Property Law & Practice, Volume 15(1), (pp 18–32).

Lee, N. (2020). Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent-product database. Technovation, Volume 96–97, 102140.

H. Ren, Y. Zhao. (2021). Technology opportunity discovery based on constructing, evaluating, and searching knowledge networks. *Technological Forecasting and Social Change*, (pp. 102196).

Bessen, J. E., & Hunt, R. M. (2007). An empirical look at software pa040307tents. Journal of Economics & Management Strategy, Volume 16(1), (pp 157–189).

Gloor, P. A. (2017). Sociometrics and Human Relationships: Analyzing Social Networks to Manage Brands, Predict Trends, and Improve Organizational Performance. Emerald Publishing Limited. DOI: 10.1108/9781787141124.

Doe, J. (2019). The impact of digital platforms on news and scientific research dissemination. Journal of Digital Information Management, Volume 17(2), (pp 34–45).

Hashimoto, D. A., Rosman, G., Rus, D., & Meireles, O. R. (2018). Artificial intelligence in surgery: Promises and perils. Annals of Surgery, Volume 268(1), (pp 70–76).

Kumar, V., & Rahman, Z. (2020). Machine learning in finance: The case of deep learning for financial markets. Journal of Financial Data Science, Volume 2(4), (pp 8–19).

Park, H., & Yoon, B. (2018). A new approach to exploring the technological opportunity discovery (TOD) using topic modeling. Technological Forecasting and Social Change, Volume 129, (pp 54–65).

Song, M., Kim, M., & Park, Y. (2017). Technological forecasting based on core author group identification and topic analysis. Expert Systems with Applications, Volume 67, (pp 126–140).

Yoon, J., & Kim, K. (2012). Trend spotting through textual data clustering and analysis for innovation planning. Expert Systems with Applications, Volume 39(10), (pp 9152–9160).

Yoon, B., Coh, B. Y., & Lee, S. (2013). Dynamic and multi-dimensional measurement of technological opportunity: Focused on the case of information and communications technology. Technological Forecasting and Social Change, Volume 80(6), (pp 1190–1207).

Zhang *et al*. (2017). Comment on the work of "Some new inequalities related to the Hermite-Hadamard inequality." Journal of Inequalities and Applications, Article 84.

Chen, L., Bae, R., Shaozheng, Q. (2018): Positive Attitude Toward Math Supports Early Academic Success: Behavioral Evidence and Neurocognitive Mechanisms. Psychological Science, Volume 29(3), (pp 390–402).

Jérôme Petigny; Christophe Ménigault *et al*. (2019). Étude économique sur l'industrie, les marchés et les déchets du plastique au Canada. Environnement et Changement climatique. https://www.publications.gc.ca/site/fra/9.871297/publication.html

A Gonçalves; G Cardeal *et al.* (2024). Sustainable Value Roadmap for the Plastics Industry. Environmental Sustainability, Volume 7(1), (pp 79–94).

A.L. Porter, S.W. Cunningham. (2004). Tech Mining: Exploiting New Technologies for Competitive Advantage. Wiley. SBN: 978-0-471-69846-3.

Y. Zhang, A.L. Porter, D. Chiavetta. (2017). Scientometrics for Tech Mining: An Introduction. In: Tech Mining: The Use of Technological Information for Decision Making, (pp 1–22).

Duriau, VJ., Reger, RK & Pfarrer, MD. (2007). A content analysis of the content analysis literature in organization studies: research themes, data sources, and methodological refinements. Organizational, Volume 10(1), (pp 5–34).

Vladimer B. Kobayashi *et al.* (2018). Text classification for organizational researchers: a tutorial. Organizational Research Methods. Sage Journal, Volume 21(3), (pp 766–799).

Wiedemann. (2013). Opening to big data: computer-assisted analysis of textual data in social sciences. Historical Social Research/Historische Sozialforschung, Volume 14, No. 2, Art. 23.

Choi *et al*. (2013). An SAO-based text-mining approach for technology roadmapping using patent information. R&D Management. https://doi.org/10.1111/j.1467-9310.2012.00702.x.

Liwei Zhang & Liu Zhihui. (2020). Research on technology prospect risk of high-tech projects based on patent analysis. PLoS ONE 15(10) : e0240050. https://doi.org/10.1371/journal.pone.0240050.

Xinyue Hu *et al*. (2024). Mapping the field: A bibliometric literature review on technology mining. Heylon, Volume 10(1), e23458.

Jinfeng Wang *et al*. (2023). Development of technology opportunity analysis based on technology landscape by extending technology elements with BERT and TRIZ. *Technological Forecasting and Social Change*, Volume194, 122341.

Ashkan Ebadi. (2022). Detecting emerging technologies and their evolution using deep learning and weak signal analysis. Technological Forecasting and Social Change, Volume 178, 121551.

Zhang Yi *et al*. (2017). Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016. Knowledge-Based Systems, Volume 133, (pp 255–268).

Alain Perez *et al*. (2018). A case study on the use of machine learning techniques for supporting technology watch. Data & Knowledge Engineering, Volume 117, (pp 239–251).

L. Kahaner. (1997). Competitive Intelligence: How to Gather Analyze and Use Information to Move Your Business to the Top. New York: Simon & Schuster. https://books.google.ca/books/about/Competitive_Intelligence.html?id=K3QfGoGSzmoC&redir_es

Armentano, M.G., Godoy, D., Campo, M., & A. Amandi. (2014). NLP-based faceted search: experience in the development of a science and technology search engine. Expert Syst. Appl., 41(6) (2014), (pp. 2886–2896).

Daniel San *et al*. (2021). A model for automated technological surveillance of web portals and social networks. Journal of Technology Management & Innovation, Volume 16(2), (pp 45-61).

Kostoff & Schaller. (2001). Science and technology roadmaps. IEEE Transactions on Engineering Management, Volume 48(2), (pp 132–143).

Xuefeng Wang *et al*. (2015). Identification of technology development trends based on subject–action–object analysis: The case of dye-sensitized solar cells. Technological Forecasting and Social Change, Volume 98, (pp 24–46).

Donghua Zhu & Porter Alan. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. Technological Forecasting and Social Change. Volume 69(5), (pp 495–506).

Yuan Zhou *et al*. (2020). Forecasting emerging technologies using data augmentation and deep learning. Open access, Volume 123, (pp 1–29).

Chau Minh, P., & Alexander, H. (2024). TopicGPT: A Prompt-based Topic Modeling Framework. arXiv:2311.01449.

Tshibanda, F., Adda, M., & Echchakoui, S. (2024) Application of machine learning in technological forecasting. *The 15th International Conference on Emerging Ubiquitous Systems and Pervasive Networks*.