UNIVERSITÉ LUMIÈRE LYON 2
MASTER 1 SCIENCES COGNITIVES

Rapport de projet

Identification des textes humains et générés par l'IA

Réalisé par :
Shenglan CHEN (2241071)

Enseignante:
Emmanuelle Reynaud

TD Informatique et programmation
2ème Semestre
Année universitaire 2024-2025

Table des matières

I. Présentation du projet

Les grands modèles de langage (LLM) basés sur l'IA transforment l'éducation et la recherche, mais soulèvent aussi des enjeux éthiques, notamment liés au plagiat. Ce projet, inspiré de l'étude de Hayawi et al. (2024), vise à distinguer les textes rédigés par l'humain de ceux générés par IA (GPT-3), à l'aide de modèles de ML et DL. Le jeu de données utilisé est disponible sur GitHub. Il comprend 28 662 résumés d'articles à propos de la pandémie de COVID-19, 50% étant générés par IA et 50% rédigés par des humains.

II. Méthodologie

2.1 Les prétraitements des données.

Avant l'extraction des caractéristiques textuelles, un nettoyage des résumés a été effectué afin d'éliminer les éléments non informatifs susceptibles de biaiser l'analyse. Ce prétraitement s'est articulé en deux étapes, désignées ici par A et B. **La phase A** a été conçue pour répondre aux besoins de l'analyse stylométrique. Elle inclut la suppression des symboles non linguistiques. Le texte est ensuite normalisé par la mise en minuscule, la suppression des chiffres et la réduction des espaces superflus. Cette version nettoyée conserve les signes de ponctuation finaux (. ! ?) , les mots fonctionnels et le lexique complet afin de permettre l'analyse fine de la richesse lexicale, de la syntaxe et de la lisibilité. **La phase B** correspond à un filtrage plus poussé, effectué à partir des textes déjà nettoyés en A, dans le but de préparer les données pour l'entraînement des modèles de classification (régression logistique, SVM, LSTM). Elle consiste principalement à retirer les signes de ponctuation finaux, les mots dits vides (stop words) i.e. qui n'apportent pas de sens, et certains termes très fréquents dans le corpus mais peu discriminants (tels que "study" ou "paper"). Ce traitement permet de réduire le bruit et de concentrer l'attention des modèles sur les termes les plus informatifs pour la tâche de discrimination IA vs humain.

2.2 Analyse stylométrique

Afin de comparer finement les productions humaines et celles générées par l'IA, une analyse stylométrique a été menée à partir des textes prétraités (version A). Cette analyse repose sur plusieurs familles d'indicateurs linguistiques, couvrant les dimensions lexicales, syntaxiques, morphosyntaxiques ainsi que la lisibilité des textes.

Caractéristiques lexicales. Le type-token ratio (TTR) est utilisé pour évaluer la diversité du vocabulaire. Il est défini comme le rapport entre le nombre de mots uniques et le nombre total de mots, sert d'indicateur de la richesse lexicale. Un TTR élevé indique une faible répétition lexicale, potentiellement révélatrice d'un style plus varié.

Caractéristiques syntaxiques. La longueur moyenne des phrases (nombre de mots par phrase) a été utilisée comme indicateur de complexité structurelle. Cette mesure permet de détecter une éventuelle tendance à produire des phrases plus longues et potentiellement plus denses en information ou en structure syntaxique.

Caractéristiques morphosyntaxiques. L'analyse morphosyntaxique a été réalisée à l'aide du modèle spaCy, permettant d'attribuer à chaque mot sa catégorie grammaticale. Deux types de mesures ont été retenus : 1) la distribution des bigrammes de catégories grammaticales (POS n-grams), permettant de détecter les enchaînements typiques propres à chaque style d'écriture ; 2) le rapport entre mots fonctionnels et mots de contenu, où les mots fonctionnels (déterminants, prépositions, conjonctions, etc.) sont opposés aux mots porteurs de sens (noms, verbes, adjectifs, adverbes). Ce ratio fournit une indication du style global.

Lisibilité et complexité. Enfin, la lisibilité a été évaluée à l'aide du score Flesch-Kincaid Grade Level, qui estime le niveau scolaire requis pour comprendre un texte. Ce score prend en compte la longueur moyenne des phrases et le nombre de syllabes par mot. Un score plus élevé traduit une complexité syntaxique et lexicale accrue. Cette mesure est particulièrement pertinente pour examiner si les textes IA présentent une structure plus formellement complexe mais pas nécessairement plus informative.

2.3 Modèles de classification

Les algorithmes de ML permettent d'extraire des caractéristiques à partir des données afin d'identifier des motifs récurrents et de construire des modèles prédictifs. Grâce à cette approche, le système peut découvrir des tendances et faire des prédictions sans intervention humaine. Dans ce projet, quatre algorithmes de ML sont utilisés :

1. **Forêt d'arbres décisionnels (RF, Random Forest)** : cet algorithme combine plusieurs arbres de décision pour établir un vote collectif, ce qui permet de mieux capter les différences subtiles entre les styles d'écriture humain et IA.
2. **Régression logistique (LR, Logistic Regression)** : elle attribue à chaque texte une probabilité d'appartenance à la classe "humain" ou "IA", ce qui en fait un bon modèle de base pour comparer la performance des autres algorithmes.
3. **Classification à vecteurs de support linéaires (Linear SVC)** : il s'agit d'une implémentation efficace du SVM linéaire, adaptée aux données de grande dimension comme les textes vectorisés en TF-IDF. Ce modèle trace une frontière linéaire optimale entre les deux types de textes, en maximisant la marge entre les exemples d'entraînement.
4. **Réseaux neuronaux Long Short-Term Memory (LSTM)** : cet algorithme de deep learning est un type avancé de réseau de neurone récurrent (RNN) conçu pour surmonter les limites des RNNs classiques. Grâce à une architecture spécifique combinant cellules mémoire, portes d'entrée, portes d'oubli et portes de sortie, le

LSTM est capable de retenir les informations essentielles sur de longues séquences et d'éliminer les données non pertinentes.

Tous les modèles ont été entraînés à partir d'une unique représentation des textes : la pondération TF-IDF (Term Frequency–Inverse Document Frequency).

III. Résultats et conclusions

3.1 Résultats de l'analyse stylométrique

Figure 1. *comparaison du type-token ratio (TTR)*

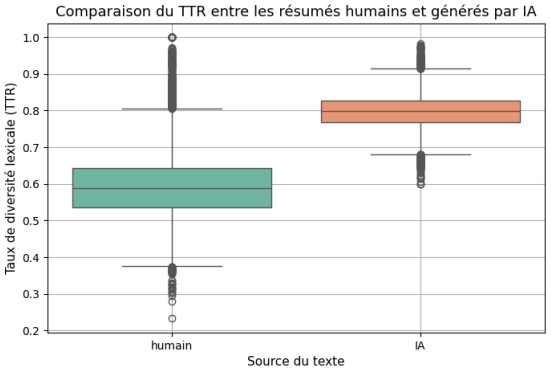
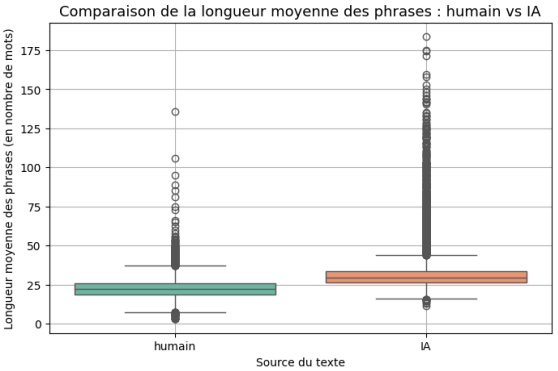


Figure 2. *comparaison du nombre de mots par phrase*



Caractéristiques lexicales

Les résumés générés par IA présentent un TTR significativement plus élevé, traduisant une diversité lexicale plus grande (Cf. Figure 1). Cela peut s'expliquer par la tendance des modèles à éviter la répétition en variant les termes ou les structures syntaxiques. À l'inverse, les textes humains montrent une répétition lexicale plus marquée, reflet de stratégies de reformulation plus limitées.

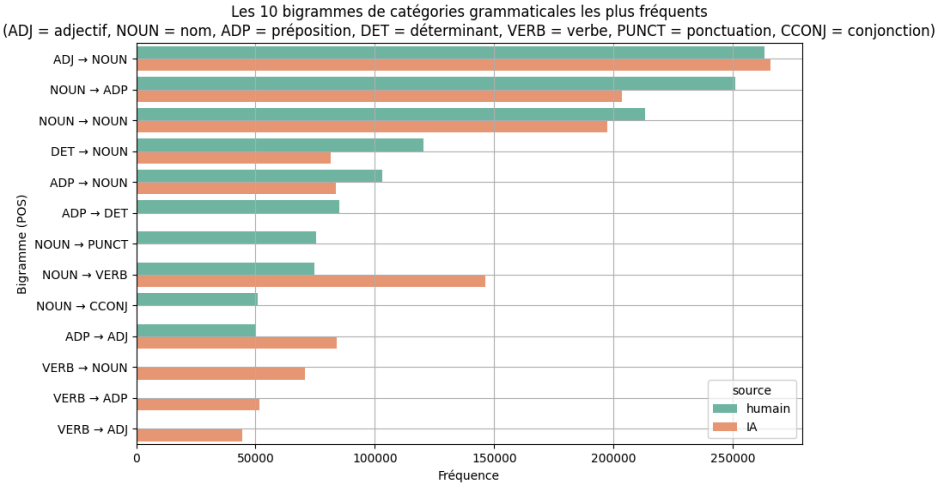
Caractéristiques syntaxiques.

La Figure 2 présente que les textes humains ont tendance à adopter des phrases plus courtes, avec une médiane de 22.56 mots par phrase. La distribution est relativement resserrée, suggérant une structure syntaxique plus régulière (SD=6.47). En revanche, les textes générés par IA présentent une tendance à adopter des phrases plus longues (M = 31.78, SD = 11.98). Ces différences traduisent que les humains privilégient généralement des phrases plus courtes et directes, tandis que les modèles d'IA ont tendance à empiler plusieurs propositions dans une même structure pour générer un texte fluide et détaillé.

Caractéristiques morphosyntaxiques

La Figure 3 compare les séquences les plus fréquentes de catégories grammaticales (POS, *part-of-speech*) dans les résumés humains et IA. Les bigrammes ADJ → NOUN (ex.: recent study), NOUN → ADP (treatment of) et NOUN → NOUN (patient population) dominent les deux sources, traduisant une forte présence de groupes nominaux et de compléments prépositionnels. Les textes générés par l'IA présentent davantage de séquences comme NOUN→VERB (model predicts) ou ADP → ADJ (under normal conditions), suggérant une tendance à produire des phrases à structure verbale plus développée. À l'inverse, les textes humains utilisent plus souvent DET → NOUN (an effect) ou ADP → NOUN (in patients), témoignant un style plus ancré dans l'énonciation concrète et la précision des référents.

Figure 3. *Les 10 bigrammes les plus fréquents (POS n-grams)*



La Figure 4 illustre la distribution du rapport entre les mots fonctionnels (tels que les déterminants, prépositions, conjonctions, pronoms, etc.) et les mots de contenu (tels que les noms, verbes, adjectifs, adverbes) dans les résumés scientifiques rédigés par des humains et par l'IA. Dans les deux cas, le rapport reste inférieur à 1, ce qui signifie que les mots de contenu sont globalement plus nombreux que les mots fonctionnels. Cependant, la médiane et l'étendue du boxplot indiquent que les résumés humains présentent une proportion relativement plus élevée de mots fonctionnels que les textes produits par l'IA. Cela reflète une tendance humaine à articuler leurs discours à l'aide de structures grammaticales comme les éléments de liaison, de coordination ou de subordination pour assurer la cohérence et la fluidité du texte. Par contre, les résumés générés par l'IA présentent une préférence pour les séquences lexicales plus denses.

Figure 4. *rapport mots fonctionnels/mots de contenu*

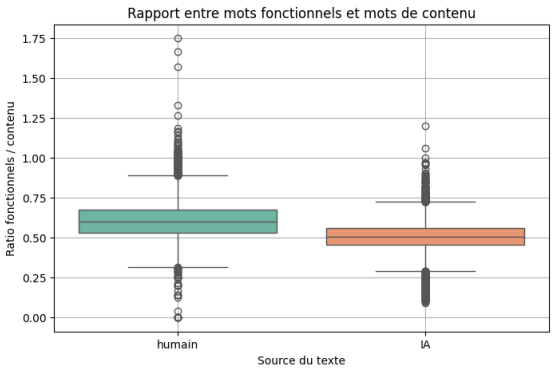
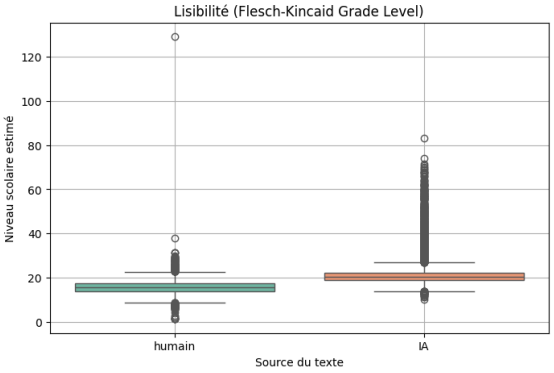


Figure 5. *Le score de Flesch-Kincaid*



Lisibilité et complexité

À l'aide du Flesch-Kincaid Grade Level, la lisibilité estimée des résumés humains et IA est présentée (Cf. Figure 5). C'est un indicateur utilisé pour estimer le niveau scolaire requis pour comprendre un texte : plus ce score est élevé, plus le texte est considéré comme complexe sur le plan syntaxique (longueur des phrases) et lexical (nombre de syllabes par mot). Généralement, le niveau attendu pour des résumés scientifiques est situé entre 12 et 18. Les résultats montrent que les résumés générés par l'IA atteignent une moyenne de 21.18 et une variabilité plus forte (SD= 5.16), alors que la difficulté à lire pour des textes humains est moins élevée et moins variable (M=15.75, SD=2.97). Cette différence indique que l'IA tend à générer des textes dont la difficulté langagière plus haute et la variabilité plus forte que celle des humains. Ce résultat est cohérent avec les tendances observées dans l'analyse du TTR et de la longueur moyenne des phrases.

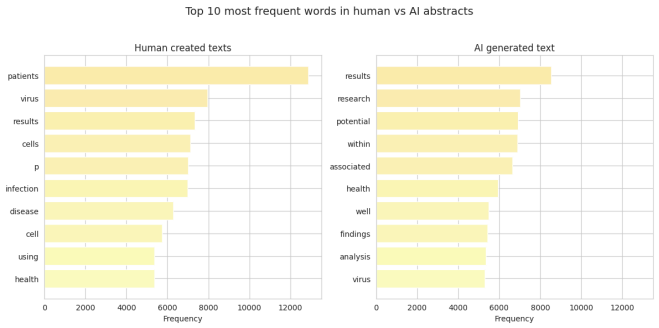
3.2 Nuage de mots-clés

Figure 6 . *Nuage de mots-clés des résumés rédigés par humain(Bleu) et des résumés générés par IA(Rouge)*

Après avoir fait la deuxième phase de nettoyage en éliminant les stopwords, afin de mieux comprendre les différences stylistiques et lexicales entre les résumés rédigés par des humains et ceux générés par l'IA, des nuages de mots-clés ont été générés pour chaque catégorie (Cf. Figure 6). La présence de mots communs aux deux catégories, tels que "results", "health" et "virus", montre un fond thématique partagé. Les résumés humains (en bleu) mettent davantage en avant des mots-clés liés à la structure classique d'un article scientifique, tels que "method", "patient" ou "model", ce qui reflète un style rédactionnel

orienté vers la précision et la description des protocoles détaillés. En revanche, les résumés générés par l'IA (en rouge) semblent privilégier des termes plus abstraits ou syntaxiquement fonctionnels comme "within", "well", "level", ce qui peut indiquer une tendance à produire un langage plus généralisant mais moins spécifique. Les 10 mots les plus fréquents des textes humain et IA (Cf. Figure 7) conviennent à cette conclusion.

Figure 7. *Les 10 mots les plus fréquents*



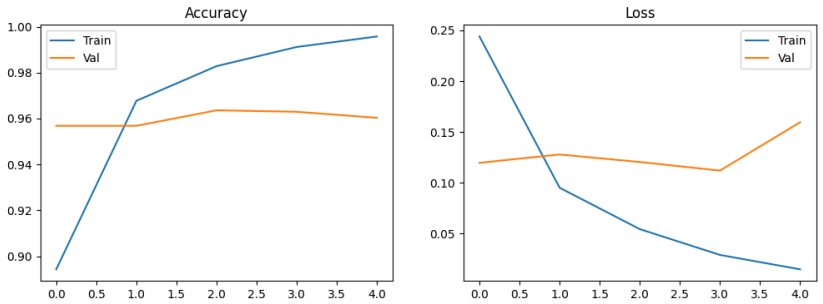
3.3 Résultats de la classification

Cette section présente les résultats des différentes méthodes d'apprentissage supervisé appliquées à la tâche de classification des résumés humains et générés par l'IA. L'entraînement des modèles a été effectué sur 80 % du corpus, soit 22 929 résumés, tandis que l'évaluation a porté sur les 20 % restants, correspondant à 5 733 résumés. Afin d'évaluer la performance des modèles de détection, plusieurs algorithmes ont été entraînés sur les données annotées (0 = humain ; 1 = IA). Les résultats sont résumés dans le tableau 1, selon quatre métriques standards : accuracy, précision, rappel et F1-score.

Table1. *Résultats de la classification*

Modèle	Classe	Accuracy	Précision	Rappel	F1-score
Random Forest	Humain	0.94	0.97	0.91	0.94
	IA		0.91	0.98	0.94
Logistic Regression	Humain	0.98	0.97	0.98	0.98
	IA		0.98	0.98	0.98
Linear SVC	Humain	0.99	0.98	0.99	0.99
	IA		0.99	0.99	0.99
LSTM	Humain	0.96 (±0.0029)	0.98	0.95	0.96
	IA		0.95	0.98	0.96

Figure 8. *Courbes d'apprentissage du modèle LSTM*



Les modèles testés présentent des performances globalement élevées, mais avec des différences selon la classe prédite. **Random Forest** atteint une précision moyenne de 0.94. Un déséquilibre entre les classes est observé : Il prédit bien la classe IA (rappel = 0.98), mais avec une précision plus faible (0.91), ce qui indique qu'il y a plus de faux positifs IA. À l'inverse, la classe Humain est prédite avec une meilleure précision (0.97), mais un rappel plus faible (0.91), ce qui suggère qu'un certain nombre de textes humains sont mal classés comme IA. **Logistic Regression** présente une performance équilibrée entre les deux classes (précision, rappel et F1-score à 0.98). Cela indique une très bonne capacité à discriminer les textes humains et IA sans biais fort envers une classe. **Linear SVC** obtient les meilleurs résultats avec des scores de 0.99 sur toutes les métriques et pour les deux classes. Il s'agit donc du modèle le plus performant et le plus stable dans ce contexte. **Le modèle LSTM** présente des performances très solides, avec une accuracy moyenne de 0.96 (±0.0029) sur l'ensemble du jeu de test. Pour la classe Humain, le modèle obtient une précision de 0.98, ce qui indique que la majorité des textes prédits comme "humain" sont effectivement écrits par des humains. Le rappel de 0.95 signifie que 5% des textes humains sont mal classés comme IA. Ce léger déséquilibre peut suggérer une certaine confusion dans les styles d'écriture très neutres ou formels. Pour la classe IA, le modèle montre un comportement symétrique : le rappel élevé (0.98) montre que la majorité des textes IA sont correctement identifiés ; la précision de 0.95 réflète que 5% des textes prédits comme IA sont en réalité humains. Bien que le modèle LSTM montre une bonne capacité d'apprentissage sur les données d'entraînement (baisse continue du loss), l'augmentation du loss de validation à partir de la 4e époque indique un début de surapprentissage (Cf. Figure 8). Cela suggère que le modèle commence à mémoriser les données au lieu de généraliser.

IV. Discussion

Dans cette étude, les caractéristiques stylistiques, lexicales, syntaxiques et morphosyntaxiques de résumés scientifiques rédigés par des humains et générés par l'intelligence artificielle sont étudiés. L'analyse stylométrique a mis en évidence des différences significatives en termes de diversité lexicale (TTR), longueur des phrases, complexité syntaxique (score Flesch-Kincaid) et usage des mots fonctionnels. Ces observations suggèrent que les modèles de langage IA tendent à produire des textes plus variés, plus longs, mais aussi plus complexes. La seconde partie de notre travail a porté sur la

classification automatique de ces résumés. Les modèles supervisés testés (Random Forest, Logistic Regression, Linear SVC, LSTM) ont tous obtenu de très bonnes performances, avec un net avantage pour le modèle Linear SVC, qui a atteint une précision quasi parfaite (0.99) sur les deux classes. Le modèle LSTM, bien qu'efficace, a montré des signes de surapprentissage à partir de la quatrième époque, soulignant la nécessité d'un ajustement plus fin pour les modèles neuronaux.

Cependant, certaines limites doivent être soulignées : D'abord, le corpus se concentre uniquement sur des résumés scientifiques liés à la COVID-19, ce qui peut restreindre la généralisation des résultats à d'autres domaines ; De plus, les textes générés par IA proviennent d'un seul modèle (GPT-3) aujourd'hui dépassé, alors que de nombreux modèles plus récents (GPT-4, Claude, Gemini) pourraient produire des textes au style différent et plus proche de l'humain ; Enfin, l'analyse stylométrique repose principalement sur des mesures superficielles (TTR, POS n-grams, lisibilité), sans prise en compte de la cohérence discursive ou du raisonnement sémantique.

Pour de futures recherches, il serait pertinent d'élargir le corpus à d'autres types de documents (emails, essais étudiants, rapports techniques) et de tester la robustesse des classifieurs face à des modèles d'IA variés et en évolution rapide. D'ailleurs, des approches d'analyse mêlant stylométrie, sémantique et détection de patterns argumentatifs pour une détection plus fine de l'origine des textes serait nécessaire.

Références

Hayawi, Shahriar, & Mathew (2024). The imitation game: Detecting human and AI-generated texts in the era of ChatGPT and BARD. Journal of Information Science, 01655515241227531.

Annexe - Code de projet

Projet_IA_HUmain