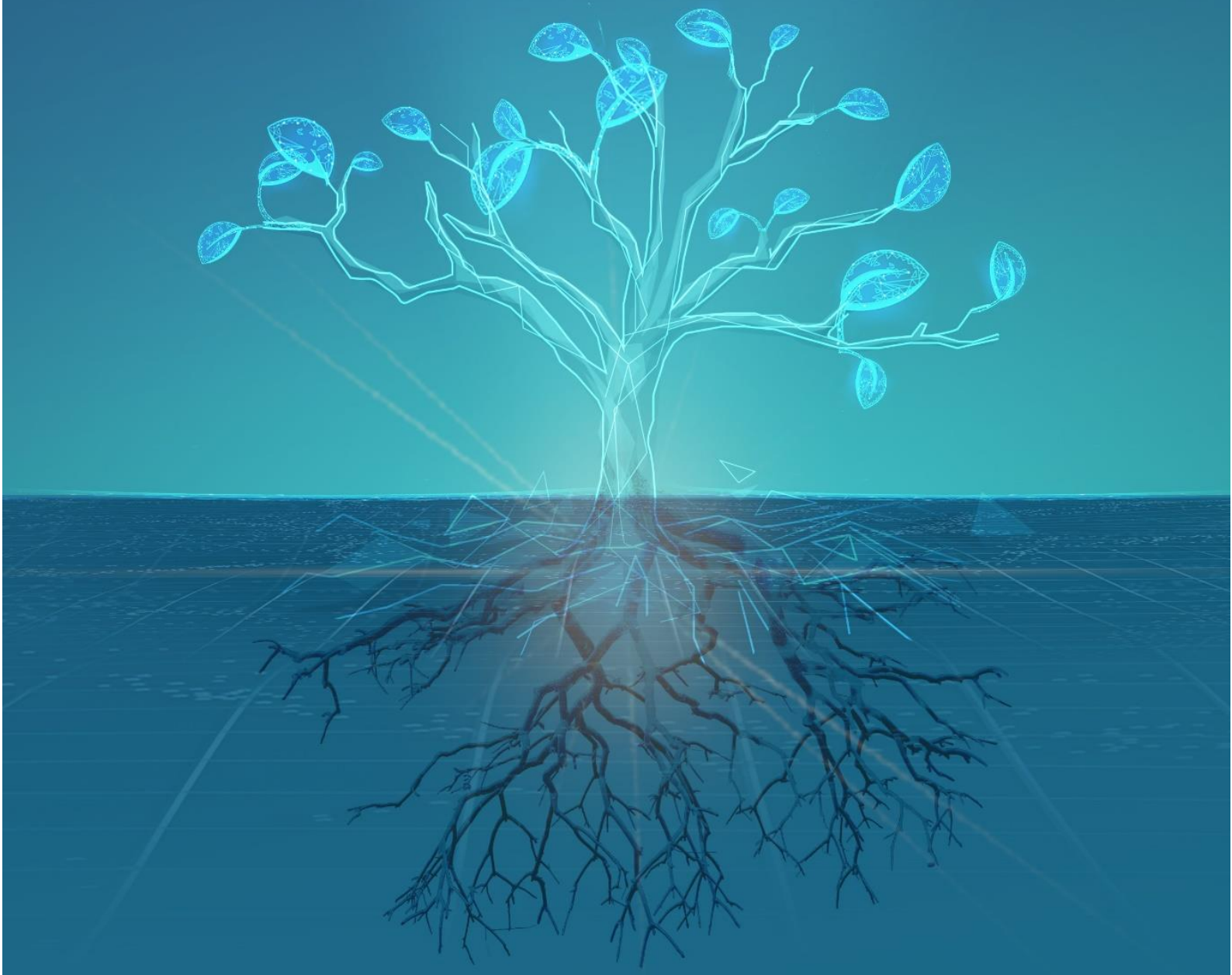
Les Cahiers de IRAFPA

Institut de Recherche et d'Action sur la Fraude et le Plagiat Académiques

Vol.1, N° 1

2023



M

Cahiers méthodologiques

Genève, 13 juillet 2023

Institut International de Recherche et d'Action sur la Fraude et le Plagiat Académiques (IRAFPA)

Site web : https://irafpa.org

Adresse postale :

IRAFPA c/o Tal Schibler, DGE Avocats

Rue Bartholoni 6

1204 Genève

Suisse

ISSN : 2813-7787

DOI : https://doi.org/10.56240/irafpa.cm.v1n1/

https://doi.org/10.56240/irafpa.cm.v1n1/rot

Détecter par stylométrie la fraude académique utilisant ChatGPT

Claude-Alain Roten, Serge Nicollerat, Lionel Pousaz, Guy Genilloud,
OrphAnalytics, Vevey (Suisse)

Mots-clefs : intelligence artificielle, ChatGPT, fraude académique, stylométrie

Résumé : Depuis la fin de l'année 2022, ChatGPT rend possible la rédaction de textes par Intelligence Artificielle. Cette IA faible est un agent conversationnel qui ne comprend ni les textes qu'il produit, ni ceux de son corpus d'entraînement. Car, pour écrire des textes crédibles, le chatbot construit des séquences de mots en choisissant les termes voisins les plus probables d'après les textes d'entraînement utilisés pour son modèle de langage GPT-3.5. Donc, par nature, ChatGPT rédige sans créativité un contenu plausible, mais pas nécessairement conforme à la réalité. Les textes ChatGPT peuvent servir à frauder dans un cadre académique : l'IA est capable de répondre à la place d'un candidat aux questions d'un examen, de rédiger un document certifiant (master, thèse…) ou d'augmenter déraisonnablement le volume des contributions d'un hyper-publiant.

L'article décrit deux approches stylométriques capables de détecter la fraude académique générée par chatbot. La rédaction d'un texte par une IA peut d'abord être mise en évidence par la comparaison de son style avec des documents authentiquement rédigés par l'auteur présumé, selon l'approche Machine Learning que nous avons développée pour détecter le ghostwriting : l'IA est soumise au même type de détection qu'un ghostwriter.

Ensuite, ChatGPT peut également être détecté comme IA indépendamment de son modèle de langage par notre approche originale Machine Learning, qui mesure le choix des mots et de leurs proches voisins : dans un texte chatbot qui préfère les voisins les plus probables, le choix est sensiblement plus restreint que dans un texte généré par un rédacteur humain.

L'article discute les conséquences de la rédaction par IA pour l'enseignement et la recherche et propose des moyens pour la détecter.

Key words: artificial intelligence, ChatGPT, academic fraud, stylometry

Abstract: Since the end of 2022, ChatGPT has made it possible to write texts using Artificial Intelligence. This weak AI is a conversational agent that understands neither the texts it produces nor those in its training corpus. In order to write credible texts, the chatbot constructs sequences of words by choosing the most likely neighbouring terms from the training texts used for its GPT-3.5 language model. So, by its very nature, ChatGPT writes content that is plausible, but not necessarily true and that is without creativity.

ChatGPT texts can be used to commit fraud in an academic context: the AI is capable of answering exam questions for a candidate, writing a certifying document (master's degree, thesis, etc.), or unreasonably increasing the volume of contributions from a hyper-publisher.

The article describes two stylometric approaches capable of detecting chatbot-generated academic fraud.

Firstly, the writing of a text by an AI can be revealed by comparing its style with documents authentically written by the presumed author by using the Machine Learning approach we have developed to detect ghostwriting: the AI is subjected to the same type of detection as a ghostwriter.

Secondly, the use of ChatGPT can also be detected independently of its language model by measuring the richness of the vocabulary: in a chatbot text that prefers the most likely neighbours, the choice is significantly more restricted than in a text generated by a human writer.

The article discusses the implications of AI-generated writing for teaching and research, and suggests ways of detecting it.

1. Introduction

Le 30 novembre 2022, la société OpenAI lançait publiquement ChatGPT-3.5, un chatbot construit sur le Grand Modèle de Langage (GML, pour Large Language Model, LLM) (OpenAI, 2022). ChatGPT répond dans une conversation aux demandes des utilisateurs, appelées « prompts ». En fournissant des réponses sous la forme d'une synthèse, l'agent conversationnel ChatGPT apparaît comme une concurrence disruptive face aux moteurs de recherche, même si, actuellement, les résultats des recherches effectuées par ChatGPT sont généralement non-sourcés. La génération automatique de textes par robots conversationnels (generative AI) profite des avancées en Intelligence Artificielle (IA) réalisées par différents acteurs : GAFAM

(Google, Meta…), startups (OpenAI appuyée par Microsoft) ainsi qu'une communauté contribuant au développement de l'IA par la création de logiciels libres (open source).

Au début de l'année 2023, ChatGPT utilise le GML GPT-3.5. L'acronyme GPT signifie Generative Pre-trained Transformer (pour Transformeur Générateur Pré-entraîné). ChatGPT se distingue des transformeurs précédents par le développement de son modèle propriétaire OpenAI et par la mise en place d'un processus contradictoire de contrôle des dérives inhérentes à tout transformeur / générateur IA. Grâce au contrôle additionnel des dérives d'une IA par GPT-3.5 (apprentissage par renforcement), OpenAI a été la première entreprise à apporter une solution aux textes erronés, voire complotistes, qui apparaissent avec l'usage des robots conversationnels tchatcheurs. Ce contrôle des dérives de langage a facilité la diffusion à large échelle de ChatGPT.

Plus précisément, ChatGPT construit un texte en choisissant une sémantique probable à partir des mots du prompt envoyé par l'utilisateur ; pour cela, il se réfère aux textes avec lesquels il a été entraîné. Concrètement, le transformeur écrit mot après mot, en choisissant le mot le plus probable dans son corpus d'entraînement. ChatGPT utilise pour cela des données statistiques complexes, rassemblées dans son GML.

Par essence, ChatGPT ne comprend donc ni les textes d'entraînement, ni les textes qu'il rédige. Il correspond actuellement à la définition d'une IA faible, incapable de réfléchir. Ses capacités cognitives sont inférieures à celles de l'intelligence humaine et à celles d'une IA forte. Pour illustrer les limites cognitives du modèle GPT-3.5 de ChatGPT, il suffit par exemple d'examiner sa stratégie au jeu d'échecs : selon les experts, si ses ouvertures académiques sont inspirées, voire plagiées des parties disponibles sur le web, ses milieux de partie se caractérisent par un déplacement aléatoire des pièces (Ft. Science4All, 2023).

Comme il construit son texte à partir des mots du prompt en utilisant une sémantique probable, le transformeur ChatGPT peut proposer un développement erroné : à notre question de savoir si Corneille a versifié une pièce de Molière, ChatGPT a répondu que Corneille avait versifié la pièce la plus célèbre de Molière *Le Cid*, alors que cette pièce, qui a fait la célébrité de Corneille, a été représentée pour la première fois en

1637, au moment où Molière (1622-1673) n'avait que quinze ans. Cet exemple montre que ChatGPT rédige sans esprit critique un texte non-sourcé au ton encyclopédique.

Un texte produit par ChatGPT, construit sur une sémantique probable, se caractérise généralement par une pensée consensuelle dans un domaine intellectuel donné. On peut donc craindre qu'il soit difficile pour cette IA de sortir du consensus. Plusieurs personnes utilisant ChatGPT pour obtenir une réponse à une question identique vont très vraisemblablement obtenir un raisonnement similaire, mais d'un phrasé différent, indétectable par les outils anti-plagiats. ChatGPT n'est clairement pas le meilleur moyen pour sortir d'un consensus intellectuel et créer de l'innovation. Utilisé sans précautions pour ses capacités intéressantes de brainstrorming ou pour faire un résumé notamment, le chatbot pourrait devenir responsable d'une perte notable de la capacité d'innovation.

En raison de ces limitations, il paraît essentiel de pouvoir contrôler l'usage abusif du transformeur en détectant ce qui est produit par l'IA, ce que ne peuvent pas faire les habituels outils de détection du plagiat. Actuellement, les détecteurs de textes rédigés par ChatGPT utilisent la connaissance des modèles GML qui ont permis la rédaction de ces mêmes textes. Par ailleurs, les fournisseurs d'agents conversationnels représentent un défi pour l'activité des moteurs de recherche parce qu'ils produisent des textes non-sourcés. Actuellement, la qualité des résultats des recherches sourcées de Google donne l'avantage au moteur de recherche, car les résultats obtenus proviennent de bases de données de qualité qui favorisent les rédactions humaines au détriment d'IA comme ChatGPT, souvent redondantes, biaisées, parfois même erronées.

De plus, la grande accessibilité de ChatGPT crée une concurrence entre les fournisseurs des moteurs de recherche et ceux des transformeurs générateurs IA. Pour une majorité d'utilisateurs, la publication des résultats de la recherche sous la forme d'un texte IA est attrayante : elle correspond à leur attente parce qu'elle est simple et immédiatement disponible. Pour éviter que les productions de textes IA ne soient déclassées car jugées de qualité médiocre voire frauduleuse, les fournisseurs de chatbot sont donc tentés de les rendre indétectables. Si cette évolution se vérifie, on aura un besoin crucial de stratégies de détection alternatives. Pour répondre à ce besoin, on propose ici deux techniques de détection stylométriques (par mesure de

style). Ces techniques sont indépendantes des modèles utilisés par les transformeurs pour rédiger les textes. Elles enrichissent les outils de détection de textes IA pour répondre aux inquiétudes formulées par les enseignants et les chercheurs.

2. La détection de ChatGPT sans connaissance des modèles de langage

2.1. La comparaison de styles pour détecter la rédaction des textes par ChatGPT

Le groupe d'experts d'OrphAnalytics a développé une méthode de comparaison des styles par apprentissage automatique (Machine Learning) qui permet de détecter si l'auteur présumé a produit lui-même le texte. On compare le style d'un échantillon de documents de référence réellement écrits par l'auteur présumé avec celui du document à authentifier. La ressemblance entre ces deux styles est une évidence forte que le document à authentifier a vraisemblablement été écrit personnellement par l'auteur présumé ; des différences de style très marquées sur une partie ou sur l'ensemble du texte seront au contraire un indice qu'il a très probablement été en partie ou intégralement plagié ou rédigé par un écrivain fantôme. L'avènement de ChatGPT pose une question importante : est-il possible de détecter une rédaction ChatGPT avec nos outils de comparaison de style capables de détecter du texte ghostwrité ou plagié ? L'approche stylométrique d'OrphAnalytics utilise des outils algorithmiques développés au cours d'une vingtaine d'expertises. Certaines avaient des enjeux judiciaires, d'autres des enjeux académiques. Pour chacune de ces expertises, les experts ont un devoir de confidentialité. Mais, en 2021, l'actuel responsable de l'enquête sur l'Affaire Grégory, le « cold case » le plus célèbre de France, a autorisé OrphAnalytics à communiquer qu'elle a délivré une expertise dans cette affaire, fondée sur des analyses stylométriques obtenus par comparaisons séquentielles. Ces résultats ont permis de déterminer qui, dans un groupe de suspects, écrit avec un style très semblable à celui des trois lettres de menace anonymes envoyées avant l'enlèvement et à celui du message de revendication du crime (OrphAnalytics, 2021). En outre, lors de la dernière élection présidentielle américaine, les experts d'OrphAnalytics se sont mobilisés pour analyser le corpus QAnon, organisation considérée comme terroriste par le FBI. Trois semaines avant les émeutes du 6 janvier 2021 au Capitole, l'équipe d'OrphAnalytics a identifié la présence de deux styles

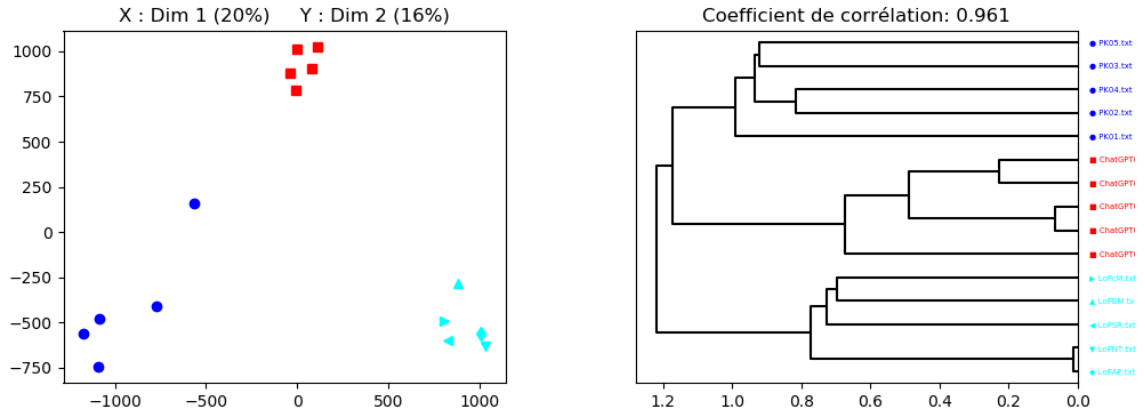
différents dans le corpus QAnon, réfutant ainsi l'hypothèse selon laquelle ce corpus, comme le suggérait la signature anonyme « Q », aurait été rédigé par une seule personne. Ces analyses stylométriques ont permis d'établir l'origine des manipulations au sein de la mouvance QAnon (OrphAnalytics, 2020 ; Gilbert, 2020). Les experts d'OrphAnlytics ont identifié par la suite deux activistes qui écrivent chacun avec un style distinct, correspondant aux deux styles observés dans le corpus QAnon. Ces résultats, corroborés par ceux du groupe de stylométrie de l'École des Chartes de Paris, ont aidé les journalistes du *New York Times* à identifier parmi un groupe de suspects les auteurs les plus probables de ce corpus (Kirkpatrick, 2022 ; OrphAnalytics, 2022). Étant donné l'impact de ChatGPT, le groupe de recherche et de développement d'OrphAnalytics a procédé à des analyses qui prouvent que la stylométrie séquentielle (appliquée sur de courtes séquences), capable d'authentifier un texte et d'en déterminer le nombre de rédacteurs (un ou plusieurs), fonctionne également avec les textes rédigés par ChatGPT (détails en légende des figures 1a et 1b). Si cette IA peut produire un texte crédible, elle est en revanche actuellement incapable de rédiger un texte respectant le style d'un auteur particulier, tel que les nombreux critères stylométriques utilisés par OrphAnalytics permettent de le définir. C'est ce que montre la comparaison entre cinq éditoriaux économiques de Paul Krugman (ronds bleus), cinq articles écrits par ChatGPT sur le thème de l'un de ces éditoriaux (carrés rouges), et cinq autres articles ChatGPT sur la Libération de Paris en 1944 écrits dans cinq styles distincts (triangles cyan) (figure 1a).

Les textes soumis à l'analyse stylométrique doivent être de taille comparable – une taille cible de 3000 caractères environ. On a établi une statistique sur l'usage de patterns de caractères (trigrammes). L'analyse multivariée en coordonnées principales PCoA (figure 1a) révèle que chaque groupe de cinq textes se caractérise par un style parfaitement distinct de celui des deux autres groupes de cinq textes.

L'analyse de clusterisation par arborescence (figure 1b) croisée avec les résultats obtenus par l'analyse multivariée confirme la répartition en trois groupes de cinq textes : le style des éditoriaux de Krugman se distingue parfaitement des deux styles de ChatGPT, utilisés par les cinq réponses aux deux prompts distincts.

La comparaison des textes / réponses à deux prompts distincts (figure 1a, clusters ChatGPT carrés rouges et triangles cyan) révèle que le style de ChatGPT en réponse

à l'un des deux prompts (e.g. rouge) est différent du style produit en réponse à l'autre prompt (e.g. cyan). Il n'y a donc pas un style unique et propre à ChatGPT. La rédaction par mots probables de ChatGPT en réponse à deux prompts différents entraîne une différence de style, comme on peut l'observer en comparant les clusters ChatGPT rouge et cyan. La dispersion plus large dans le cluster bleu des textes humains, révèle également la créativité bien plus grande dans les textes de Paul Krugman par rapport aux textes de l'IA (figure 1a). La plus grande créativité de Paul Krugman est confirmée par la très grande longueur des branches du cluster bleu dans l'arbre de la figure 1b.



Figures 1 : En ronds bleus, cinq éditoriaux de Paul Krugman du New York Times, en carrés rouges, cinq articles économiques rédigés par ChatGPT sur les mots-clés d'un des cinq éditoriaux de Paul Krugman publiés, en triangles cyan, cinq articles de ChatGPT décrivant la libération de Paris en 1944 comme si l'événement venait de se dérouler en utilisant les styles suivants très différents : Associated Press, New York Times, Corman MacCarthy, Salman Rushdie, Livre des Mormons. Taille cible moyenne des textes : environ 3000 caractères.

Figure 1a, à gauche : Figure bidimensionnelle des deux dimensions de variance maximale obtenues après l'analyse multivariée PCoA des distances Manhattan séparant les quinze textes positionnés dans un espace défini par les fréquences de trigrammes de caractères.

Figure 1b, à droite : Clusterisation par arborescence moyennant les distances angulaires cos thêta séparant les points obtenus après la PCoA de la fig. 1a, dans les dimensions résultantes couvrant le 90% de la variance cumulée.

Les figures 1a et 1b révèlent deux styles de ChatGPT (clusters rouge et cyan). Ces deux styles sont distincts de celui de Paul Krugman (cluster bleu). Les variations du style rend la fraude par rédaction IA parfaitement détectable par les techniques d'analyse stylométrique. Par exemple, si un auteur a rédigé frauduleusement des parties d'un texte à l'aide de ChatGPT, l'analyse stylométrique par comparaison de styles, indépendante des modèles IA, détectera deux styles se manifestant dans les deux clusters différents de l'humain et de l'IA (respectivement clusters bleu et rouge

de la figure 1a). La rédaction de tout ou de parties d'un texte par une IA est donc détectable grâce à l'authentification stylométrique par comparaison de style, et ce par nature, quel que soit le générateur IA utilisé.

2.2. La détection de ChatGPT par mesure de degré de liberté de choix sémantique

La stratégie de rédaction de ChatGPT – choisir mot après mot celui qui est le plus probable – permet une stratégie de détection qui ne nécessite pas la connaissance du corpus d'entraînement du chatbot ou de son modèle de langage.

La stratégie de sélection du mot probable permet au chatbot d'écrire du texte crédible sans comprendre ni les textes d'entraînement ni ceux qu'il rédige. ChatGPT produit majoritairement des textes de styles similaires à ceux des corpus d'entraînement ; en revanche, un rédacteur humain est principalement guidé par ses objectifs de rédaction, le style n'étant qu'une conséquence. Une rédaction humaine se caractérise donc par un choix de mots plus libre et plus riche que celui de ChatGPT. Une mesure de la richesse du vocabulaire et de la combinaison des mots peut donc permettre d'identifier une source IA par différence avec une source humaine.

La figure 2 illustre comment détecter les textes ChatGPT selon cette approche. Après traitement par Machine Learning, les contraintes de vocabulaire, c'est-à-dire les répétitions pondérées en fonction de la longueur des textes, sont représentées par des colonnes de couleur dont la longueur est proportionnelle à ces contraintes : au centre en bleu, les cinq articles économiques de l'éditorialiste Krugman ; à gauche en rouge, les cinq textes économiques rédigés par ChatGPT selon le thème de l'un des cinq articles de Krugman ; et à droite en cyan, les cinq récits de ChatGPT décrivant la Libération de Paris de 1944 avec cinq styles distincts.

Dans cet exemple analysant 15 articles d'une taille cible de 3000 signes (ce qui correspond à un peu moins d'une pleine page standard de texte MS-Word, soit environ 500 mots, ou jetons pour tokens), le choix sémantique des cinq textes de Paul Krugman paraît en moyenne au moins trois fois moins contraint (barres bleues courtes) que celui des textes économiques et historiques de ChatGPT (figure 2, barres rouges et cyan plus longues). L'exemple de la figure 2 illustre ainsi la différence significative de degré de liberté dans le choix des mots : contrainte sémantique trois fois moins grande chez un rédacteur humain que pour ChatGPT.

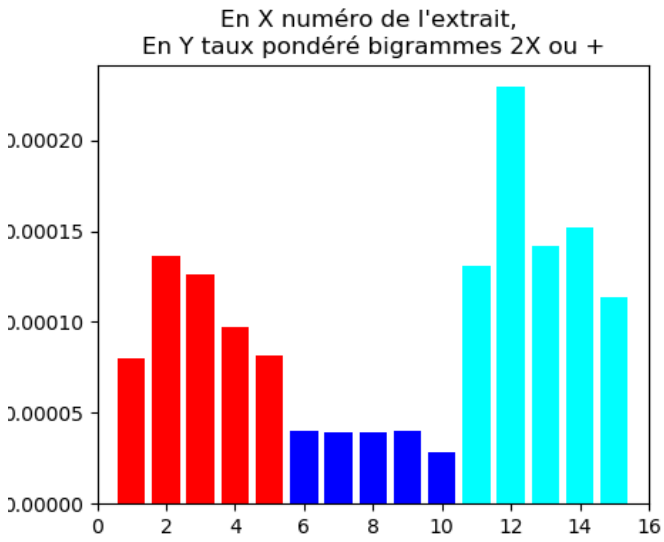


Figure 2 : Liberté de choix de mots par analyse des bigrammes fréquents de mots pondérée par la longueur des textes. Les textes et leur couleur correspondent à ceux de la figure 1.

La détection ChatGPT par comparaison de richesse de vocabulaire fonctionne sur des textes d'une taille cible minimale d'une demi-page, avec une précision estimée égale ou supérieure à 90%. Des mesures de quantification statistique sont en cours d'évaluation pour faciliter l'analyse du contenu sémantique.

La détection d'IA par mesure de la richesse du vocabulaire utilisé est par ailleurs bien plus rapide que celle que l'on peut obtenir avec d'autres détecteurs du robot conversationnel ChatGPT, car elle n'utilise que le texte questionné. De plus, cette analyse de richesse du vocabulaire est immédiatement applicable dans d'autres langues pour la comparaison de textes en masse.

3. Conclusion

Les exemples discutés révèlent que la stylométrie par comparaison de styles et par mesure de richesse de vocabulaire sont capables toutes les deux de détecter si un texte est rédigé par un humain ou s'il a été rédigé en partie ou en totalité par ChatGPT. Pour détecter l'écriture fantôme, la mesure par comparaison de style paraît la plus robuste au vu de l'évolution rapide des générateurs IA, le ghostwriter étant en l'occurrence une IA du type ChatGPT. Cette mesure nécessite simplement d'avoir à disposition d'autres textes authentiquement écrits par l'auteur présumé, afin de permettre une analyse comparative (OrphAnalytics, 2023).

La détection de rédaction par IA indépendante des modèles de transformeurs est rapide et elle permet d'identifier les usages incontrôlés et potentiellement frauduleux d'une IA dans les travaux académiques. D'autres indicateurs peuvent être associés à la mesure de richesse de vocabulaire si les fraudeurs adoptent des contre-mesures destinées à augmenter la richesse sémantique. Des algorithmes mesurant les constructions de phrases et leur agencement sont en phase d'évaluation.

Les deux outils stylométriques capables de détecter les écritures fantômes, qu'elles soient humaines ou générées par IA, sont de nature à compléter utilement les outils de détection du plagiat, qui sont devenus inopérants pour les textes générés par IA. Compilatio, le principal fournisseur de logiciels anti-plagiat des institutions francophones, et OrphAnalytics travaillent actuellement à intégrer aux outils déjà existants et fonctionnels de détection du plagiat les moyens de détecter l'écriture fantôme.

OrphAnalytics collabore également avec l'École des Sciences Criminelles de l'Université de Lausanne pour quantifier les évidences apportées par les deux types de détection stylométrique de ChatGPT – la comparaison de style et la mesure de richesse sémantique. D'autres approches stylométriques sont en cours de développement pour répondre aux défis soulevés par les nouvelles versions de ChatGPT.

Toute utilisation de ChatGPT n'est pas frauduleuse, l'IA servant à merveille à brainstormer ou à résumer de longs documents. En revanche, il faut pouvoir détecter l'usage du chatbot pour s'assurer qu'il fait (ou non) l'objet d'un contrôle humain, garant de l'authenticité des travaux et des publications.

Bibliographie

Ft. Science4All. (2023, 8 janvier). De quoi ChatGPT est-il VRAIMENT capable ? Consulté le 22 juin 2022 sur https://www.youtube.com/watch?v=R2fjRbc9Sa0

Gilbert, D. (2020, 16 décembre). QAnon's Mysterious Leader 'Q' Is Actually Multiple People, consulté le 22 juin 2023 sur https://www.vice.com/en/article/jgqj7x/qanons-mysterious-leader-q-is-actually-multiple-people

OpenAI. (2022, 30 novembre). Introducing ChatGPT. Consulté le 22 juin 2023 sur https://openai.com/blog/chatgpt/

Kirkpatrick (D. D.). (2022, 19 février). Who Is Behind QAnon ? Linguistic Detectives Find Fingerprints. Consulté le 22 juin 2023 sur https://www.nytimes.com/2022/02/19/technology/qanon-messages-authors.html

OrphAnalytics. (2020, 15 décembre). QAnon serait rédigé par deux personnes différentes, comme le montre l'analyse par machine learning. Consulté le 22 juin 2023 sur https://www.orphanalytics.com/fr/news/pressrelease20201215

OrphAnalytics. (2021, 24 avril). Communiqué de OrphAnalytics SA concernant l'affaire Grégory, consulté le 22 juin 2023 sur https://www.orphanalytics.com/fr/news/statement-2021-04

OrphAnalytics. (2022, 19 février). La linguistique computationnelle jette un nouvel éclairage sur l'identité de QAnon. Consulté le 22 juin 2023 sur https://www.orphanalytics.com/fr/news/whitepaper202201

OrphAnalytics. (2023, 27 février). Détection de la production de textes ChatGPT sans connaissance du modèle de langage utilisé. Consulté le 22 juin 2023 sur https://www.orphanalytics.com/fr/news/chatgpt-oa1