Fiche synthèse

Détecter par stylométrie la fraude académique utilisant ChatGPT

Claude-Alain Roten, Serge Nicollerat, Lionel Pousaz, Guy Genilloud

OrphAnalytics, Vevey (Suisse)

Article publié dans *Les Cahiers méthodologiques de l'IRAFPA*, IRAFPA, Vol. 1, N°1, juillet 2023.

Cet article est destiné aux enseignants universitaires, afin de leur présenter des stratégies de contrôle pour l'outil rédactionnel ChatGPT. Leur rôle est de prévenir les risques liés à l'utilisation de l'intelligence artificielle (IA) : textes plausibles, mais sans fondement, et susceptibles d'être détectés d'avoir été rédigés par l'IA. Afin de soulever les enjeux réglementaires, voire juridiques, de la rédaction par ChatGPT, l'article se focalise d'abord sur le fonctionnement de cette IA pour déterminer les contraintes de la rédaction ChatGPT afin de les exploiter ensuite, dans deux solutions stylométriques de détection de rédaction par ChatGPT, indépendantes des modèles de langage.

ChatGPT est un outil utile pour la rédaction professionnelle (pour esquisser un texte ou résumer un document). Bien utilisé, il est un excellent outil de recherche. Il peut écrire tout ou partie d'un document par sa capacité de rédaction, mais ChatGPT peut servir à la fraude. Dans un cadre d'examen ou d'évaluation par certification, il peut être utilisé de façon abusive comme un ghostwriter, un écrivain fantôme. En outre un texte IA sans source, au ton encyclopédique, peut créer des hallucinations en décalage avec la réalité.

Par son écriture sans copie, ChatGPT échappe aux détections de plagiat. Or, la détection de textes rédigés par ChatGPT devient un besoin pressant. Les détecteurs IA aujourd'hui disponibles sont développés à partir des modèles de langage qui ont justement servi au développement de la rédaction de textes par IA. La rédaction des textes IA peut donc être manipulée par les fournisseurs de rédaction IA pour rendre ces textes non détectables.

Deux solutions de détection indépendantes des modèles de langage de ChatGPT sont proposées dans cet article afin d'être garantes de l'intégrité. Elles ont été développées dans le cadre de l'entreprise OrphAnalytics à laquelle sont rattachés les auteurs de cet article.

1/ Nous avons développé, avant ChatGPT, un outil de détection de fraude capable de mettre en évidence du ghostwriting, c'est-à-dire du texte écrit par une personne autre que le candidat. Cet outil s'avère également capable de détecter la rédaction par ChatGPT de tout ou partie d'un document certifiant, car avec ces outils de stylométrie, l'écriture de ChatGPT se comporte comme celle d'un ghostwriter. Ces

analyses de comparaison de style permettent donc de mesurer si le signataire d'un document est le réel rédacteur de ce document.

2/ Une approche de mesure de richesse de vocabulaire permet de s'assurer de façon crédible si un texte a été rédigé par un agent conversationnel comme ChatGPT, car une IA écrit des textes moins riches en vocabulaire et marqués de répétition que ceux d'êtres humains.

Afin de répondre aux enjeux sociétaux, tels que la fraude à large échelle ou la tentation d'utiliser ChatGPT comme une aide aux examens ou à l'écriture des documents certifiants, l'article s'attache aux besoins essentiels de détecter pour réguler, et non pas pour sanctionner. Puisque la détection stylométrique de ChatGPT est indépendante des modèles de langage et de la langue des textes, nous pensons que notre contribution est susceptible d'apporter une solution de contrôle d'intégrité au service du respect des bonnes pratiques.

Pour appréhender la révolution IA avec un prisme positif, constatons qu'encadrés, les étudiants peuvent apprendre différemment en utilisant ce nouvel outil de recherche. L'usage contrôlé de ChatGPT devrait réduire sensiblement le risque de perte d'innovation dans les institutions académiques.