UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

RAPPORT FINAL

PRÉSENTÉ À

KHOURY, RAPHAËL (Superviseur)

EL GUEMHIOUI, KARIM (Coordonnateur)

PROJET

« ANALYSE DU BIAIS DANS LE LANGAGE PRODUIT PAR DES GRANDS MODÈLES DE LANGAGE »

PAR

HECHUN OUYANG (OUYH85020302)

COURS

PROJET SYNTHÈSE

(INF4173-SO) - AUTOMNE 2024

16/12/2024

Table de matières

Remerciements

Je veux remercier le Professeur Raphaël Khoury, pour ses conseils, sa patience, et sa confiance en moi, particulièrement au début du projet, et sa volonté d'accepter ce projet de recherche sachant que je suis en échange international, à l'Université de Jaén en Espagne. Je veux également remercier le directeur du module d'informatique Karim El Guemhioui pour ses réponses assidues à mes questions par rapport à ce projet et son support pour la possibilité de faire ce projet de recherche à distance. Finalement, je veux remercier mon collègue Yamine Ibrahima pour ses conseils concernant la rédaction et le déroulement du projet.

Introduction

Contexte

Dans le cadre de l'analyse et l'écriture de textes, les grands modèles de langage sont souvent utilisés pour synthétiser, rédiger ou reformuler du contenu. L'utilisateur doit pouvoir faire confiance aux modèles de langage pour bien représenter les idées du texte. Ils exercent alors une influence majeure sur l'écriture contemporaine. Les mots qu'ils décident d'utiliser pour représenter une idée quelconque peuvent introduire un biais inconnu à l'utilisateur.

Quelques statistiques d'utilisation de l'intelligence artificielle :

- Le modèle de langage le plus populaire est ChatGPT. ChatGPT fait partie des 20 sites web les plus visités en juillet 2024, avec plus de 3 milliards de visites.[1]
- Selon Statista, l'usage le plus commun de ChatGPT par des employés aux États-Unis en 2023 était d'écrire du contenu.[2]
- Selon une estimation, environ 10 % des articles provenant des entreprises Fortune 500 sont potentiellement générés avec de l'intelligence artificielle.[3]
- D'après certaines analyses, plus de la moitié des étudiants universitaires ont déjà utilisé de l'IA dans leurs travaux ou dans leurs examens.[4]

Ces données illustrent à quel point l'intelligence artificielle influence déjà profondément la production de contenu aujourd'hui.

---

[1] Web Archive. "Trending Websites - Global." *SEMrush*, 12 septembre 2024. https://web.archive.org/web/20240912045747/https://www.semrush.com/trending-websites/global/all.
[2] Statista. "Principaux usages de ChatGPT par les employés américains." *Statista*, consulté le 20 septembre 2024. https://www.statista.com/statistics/1441294/top-uses-of-ChatGPT-by-us-employees/.
[3] Originality.AI. "Le contenu généré par IA domine certains sites du Fortune 500." *Blog Originality.AI*, consulté le 20 septembre 2024. https://originality.ai/blog/ai-generated-content-dominates-on-some-fortune-500-sites.
[4] BestColleges. "La majorité des étudiants ont utilisé l'IA : enquête." *BestColleges*, consulté le 20 septembre 2024. https://www.bestcolleges.com/research/most-college-students-have-used-ai-survey/.

Problème

La quantification des biais introduits dans la génération de contenu est un défi bien documenté. Des études antérieures suggèrent que ChatGPT présente un biais religieux contre l'Islam et présente un biais de genre. ChatGPT associe souvent l'Islam avec la violence et recommande des livres comme *Harry Potter* principalement aux garçons.[5]

Objectif

L'objectif principal du projet est de poursuivre la recherche déjà faite concernant les biais des modèles de langage en analysant l'objectivité des termes utilisés par des textes produits par l'intelligence artificielle. Plus précisément, cette étude vise à déterminer si ChatGPT emploie un langage émotionnellement chargé ou biaisé en accord avec l'information donnée dans la requête, ou bien s'il modifie le ton lors de la génération de contenu en utilisant un langage plus positif ou négatif.

Travail réalisé

Méthodologie

La méthodologie consiste à comparer le score sentimental des articles générés avec ChatGPT à celui des articles originaux. On génère les articles en se basant sur des résumés sous forme de puces des articles originaux. Les puces sont aussi générées par ChatGPT. Les articles originaux font partie d'une base de données ouverte. Les scores sentimentaux sont extraits à l'aide d'un algorithme d'analyse sentimental intitulé Afinn. Le score sentimental extrait est ensuite analysé avec des méthodes statistiques et une étude de cas.

[5] Babaei, G., Banks, D., Bosone, C., Giudici, P., & Shan, Y. (2024). Is ChatGPT More Biased Than You? *Harvard Data Science Review*, *6*(3). https://doi.org/10.1162/99608f92.2781452d

Premièrement, on trouve une base de données d'articles provenant d'un site web comme Kaggle. Idéalement les articles sont divisés en fonction du biais politique ou, dans ce cas, divisés en fonction de leur crédibilité (fausses nouvelles ou vraies nouvelles), pour faciliter l'analyse. La base de données est nettoyée en retirant les articles trop courts (moins de 100 mots, pour garantir leur représentativité) et ceux dépassant 1200 mots, en raison des limites de génération de ChatGPT pour des articles excédant 1000 mots.

Ensuite, ChatGPT génère des puces qui résument l'article. Une fois les puces produites, un nouvel article est généré par ChatGPT en utilisant uniquement les puces comme base, sans connaissance de l'article original. On automatise ce processus en utilisant l'interface de programmation d'application (API) de ChatGPT. On utilise le modèle de GPT 4o-mini.

L'API de ChatGPT offre la possibilité d'ajuster un paramètre de température. Ce paramètre varie entre [0, 1]. En théorie, une température plus basse (0) génère des réponses avec un langage plus neutre. Une température plus élevée (1) favorise des réponses employant un langage plus créatif et expressif.[6] On génère des articles avec trois températures distinctes : 0.0, 0.5, 1.0.

Pour récapituler, pour chaque article original, on génère trois articles différents, un pour chaque température (0.0, 0.5, 1.0). Ensuite, on répète ce processus pour chaque température 5 fois afin de faire la moyenne des scores sentimentaux générés pour avoir des données moins influencées par les extrêmes.

Voici le réglage utilisé pour générer les puces :

- System Content (contexte et instructions donnés pour l'API) : « You are a professional bullet point summarizer. »
- User Content (« prompt » pour ChatGPT) : « Summarize the following news article into bullet points. Rely strictly on the provided text, without including external information.:\n\n{article_text} »

[6] Davis, J., Van Bulck, L., Durieux, B. N., & Lindvall, C. (2024). The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research. JMIR human factors, 11, e53559. https://doi.org/10.2196/53559

Voici le réglage utilisé pour générer les articles :

- System Content (contexte et instructions donnés pour l'API) : « You are a news article writer. »
- User Content (« prompt » pour ChatGPT) : « Your boss has asked you to write a news article of about {nb_words} words long, based only on the bullet points, without referencing or including any information outside of the bullet points.\n\n{bullet_points} »

Le score sentimental des articles originaux est comparé à celui des articles générés par ChatGPT. Le score sentimental représente si l'article contient du langage positif ou négatif. Par exemple, le mot « happy » se voit attribué un score sentimental de +3.0. Tandis que, le mot « sad » reçoit un score sentimental de -2.0.

Cette expérience est répétée plusieurs fois pour chaque article original. On peut ensuite agréger les scores sentimentaux générés.

En analysant les scores sentimentaux et les sujets abordés, on peut observer l'absence ou non de biais dans le langage utilisé par ChatGPT. On a choisi Afinn comme algorithme/dictionnaire d'analyse du score sentimental. Afinn donne un score précis pour chaque mot. Plusieurs autres algorithmes ou lexiques d'objectivité, par exemple, celui du CNRC, ou MPQA de l'Université de Pittsburgh, indiquent seulement si un mot est positif (1), négatif (-1), ou neutre (0), mais n'inclut pas de nuance. Tous les mots négatifs sont considérés comme autant négatifs entre eux.

Avec Afinn, on peut également facilement analyser des articles entiers et sortir des données intéressantes tel le score sentimental, le score sentimental moyen par mot, le nombre de mots, etc. Dans le cadre de cette recherche, un langage positif signifie un lexique avec un score sentimental élevé et un langage négatif signifie un lexique avec un score sentimental bas.

La banque d'articles sélectionnée pour cette analyse est la suivante :

- https://www.kaggle.com/datasets/mdepak/fakenewsnet?resource=download

Cette base de données est une collection d'articles que la compagnie Buzzfeed a classifié comme « fake news » et « real news ». Les articles analysés ne sont pas écrits par Buzzfeed. Pour le restant de l'article, on référencera ces articles comme « fausses nouvelles », « vraies nouvelles », « faux articles » et « vrais articles ». Au total, 85 fausses nouvelles ont été analysées et 77 vraies nouvelles. Des 85 fausses nouvelles, 14 proviennent de sources avec un important biais politique de gauche et 64 de droite. Des 77 vraies nouvelles, 9 proviennent de la gauche et 14 de la droite.

Voici d'autres banques d'articles potentielles qui pourraient enrichir les analyses futures (liste non-exhaustive) :

- https://www.kaggle.com/datasets/emirslspr/israel-hamas-conflict-news-dataset
- https://www.kaggle.com/datasets/stevenpeutz/misinformation-fake-news-text-dataset-79k

Voici les étapes suivies :

1. Identifier une base de données d'articles. La partie BuzzFeedNews dans la base de données FakeNewsNet a été sélectionné pour cette étude.
2. En utilisant l'API de ChatGPT, on a généré des puces pour chaque article et pour chaque température de liberté de ChatGPT (0.0, 0.5, 1.0).
3. En utilisant l'API de ChatGPT, les articles ont ensuite été générés à partir des puces produits à l'étape précédente.
4. Refaire les étapes 2 et 3 cinq fois.
5. En utilisant l'algorithme Afinn, calculer le score sentimental pour chaque article original ainsi que les articles générés.
6. Traiter les données et ressortir la médiane, la moyenne, le score sentimental par mot, la différence entre le score sentimental original et celui généré, etc.
7. Identifier les cas extrêmes où le score sentimental généré diffère grandement du score sentimental original.
8. Les articles ont été regroupés en sous-catégories selon leur affiliation politique, notamment à gauche ou à droite du spectre.
9. Les données ont été soumises à une analyse statistique approfondie, suivie d'une interprétation des résultats.

Résultats

Les articles générés par ChatGPT présentent, en moyenne, un langage plus positif que celui des articles originaux

Analyse du score sentimental dans les articles originaux et générés

Indépendamment du type de nouvelles (fausses ou vraies) ou de leur position sur le spectre politique, ChatGPT génère des articles comportant des mots plus positifs.

Tableau 1: Moyennes des scores sentimentaux pour les fausses nouvelles (85 articles)

Fausses nouvelles	Original	Température ChatGPT 0.0	Température ChatGPT 0.5	Température ChatGPT 1.0	Généré
Score sentimental par mot (moyenne)	-0.02687	-0.01976	-0.01827	-0.01991	-0.01905
Score sentimental par article (moyenne)	-14.1059	-9.21647	-8.93412	-9.93412	-9.36157

Tableau 2: Moyennes des scores sentimentaux par mot pour les vraies nouvelles (77 articles)

Vraies nouvelles	Original	Température ChatGPT 0.0	Température ChatGPT 0.5	Température ChatGPT 1.0	Généré
Score sentimental par mot (moyenne)	-0.01055	0.002058	0.001416	0.002063	0.001846
Score sentimental par article (moyenne)	-6.75325	-0.04156	-0.85714	-0.91429	-0.60433

Le score sentimental par article peut être négatif tandis que le score sentimental par mot est positif en raison de la pondération égale attribuée à chaque article dans les calculs.

Prenons par exemple deux articles : article alpha de 1000 mots avec un score sentimental de 200 et article bravo de 100 mots avec un score sentimental de -100. Le score sentimental par mot serait calculé avec la formule :

$$score\ sentimental\ par\ mot = \frac{\sum_{j=1}^{M} w_j}{M}$$

Où :

- M = nombre de mots total dans l'article
- w_j = score sentimental du j-ème mot.

La moyenne des scores sentimentaux par mot est déterminée par cette formule :

$$moyenne\ des\ scores\ sentimentaux\ par\ mot = \frac{\sum_{i=1}^{N} score\ sentimental\ par\ mot}{N}$$

Où :

- N = nombre d'articles

Dans l'exemple défini précédemment, on aurait alors pour l'article alpha un score sentimental par mot de 0.2, tandis que pour l'article bravo on aurait un score sentimental par mot de -1. Cependant, la moyenne des scores sentimentaux par mot serait :

$$\frac{0.2 - 1}{2} = -0.4$$

Cette imprécision est causée parce qu'on utilise la fonction AVERAGE dans Excel pour trouver la moyenne des scores sentimentaux.

Cependant, le score sentimental par article serait de :

$$\frac{200 - 100}{2} = 50$$

On a alors une situation où le score sentimental par mot est négatif tandis que le score sentimental par article est positif.

On cherche à quantifier comment plus positif les articles générés sont par rapport aux articles originaux. On peut trouver la différence d'objectivité relatif avec la formule :

$$Différence\ d'objectivité\ relatif\ (pourcentage) = \frac{(S_{gen} - S_{orig})}{|S_{orig}|}\ x\ 100$$

Où :

- S_{gen} = score sentimental généré
- S_{orig} = score sentimental original

On peut aussi trouver la différence d'objectivité absolu avec la formule :

$$Différence\ d'objectivité\ absolu = S_{gen} - S_{orig}$$

En appliquant ces formules, on obtient les résultats suivants :

Tableau 3: Différence d'objectivité des articles générés par ChatGPT avec les articles originaux

Type de nouvelles	Différence d'objectivité relatif (%)	Différence d'objectivité absolu
Fausses nouvelles (par mot)	29.11	0.00782
Fausses nouvelles (par article)	33.63	4.744
Vraies nouvelles (par mot)	117.5	0.012396
Vraies nouvelles (par article)	91.05	6.14892

La différence d'objectivité relatif est au-dessus de 100% quand le score sentimental d'un article original est négatif, tandis que le score sentimental généré est positif. En d'autres termes, quand on change de positif à négatif ou vis-versa. Les résultats montrent que, globalement, les articles générés par ChatGPT adoptent un langage plus positif que les articles originaux. Cette tendance est particulièrement marquée pour les articles initialement classifiés comme des vraies nouvelles.

Ceci correspond aux boîtes à moustaches représentant les données :

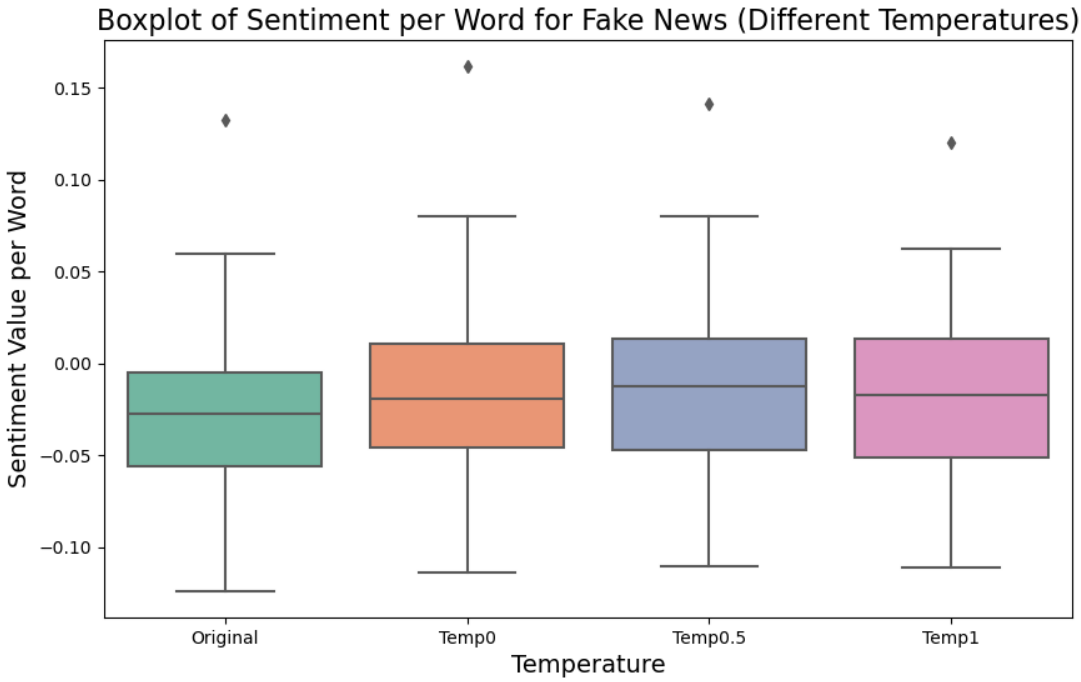

Figure 1: Boîte à moustaches du score sentimentale par mot pour les fausses nouvelles, regroupée par température.

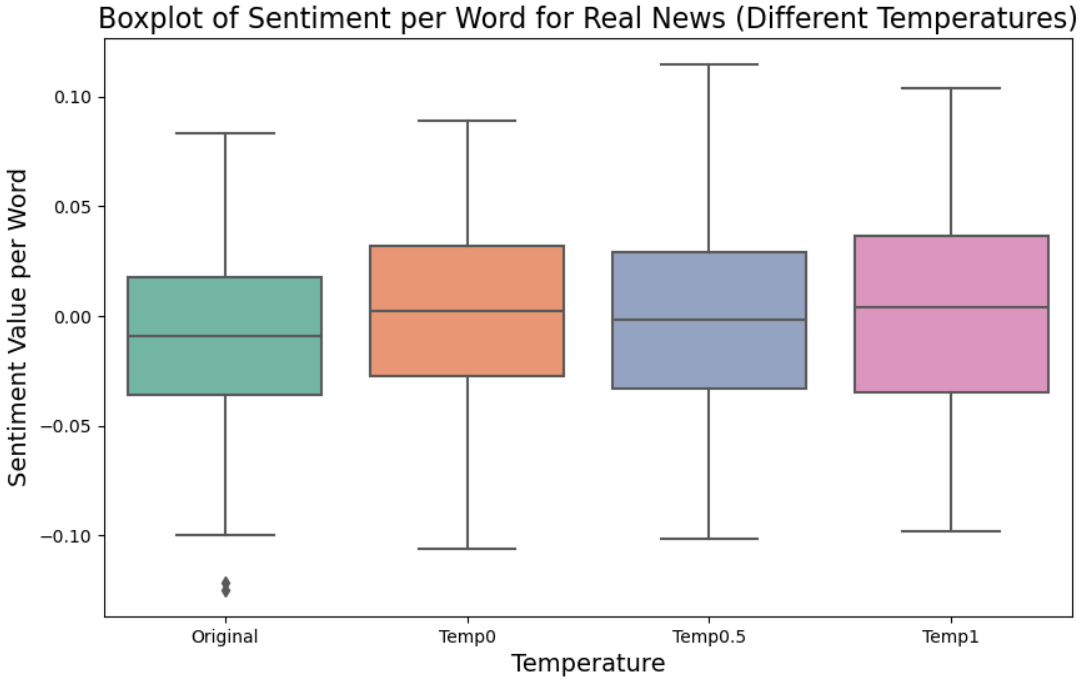

Figure 2: Boîte à moustaches du score sentimental par mot pour les vraies nouvelles, regroupée par température.

À l'aide des boîtes à moustaches, on peut voir que les scores sentimentaux des articles originaux sont généralement plus bas que ceux des articles générés.

Le score sentimental des articles suit une tendance normale

Les graphiques suivants ainsi que le test de normalité Shapiro-Wilk permettent d'affirmer que le score sentimental par mot pour les articles, soit originaux ou générés, suivent une distribution normale.

Articles faux

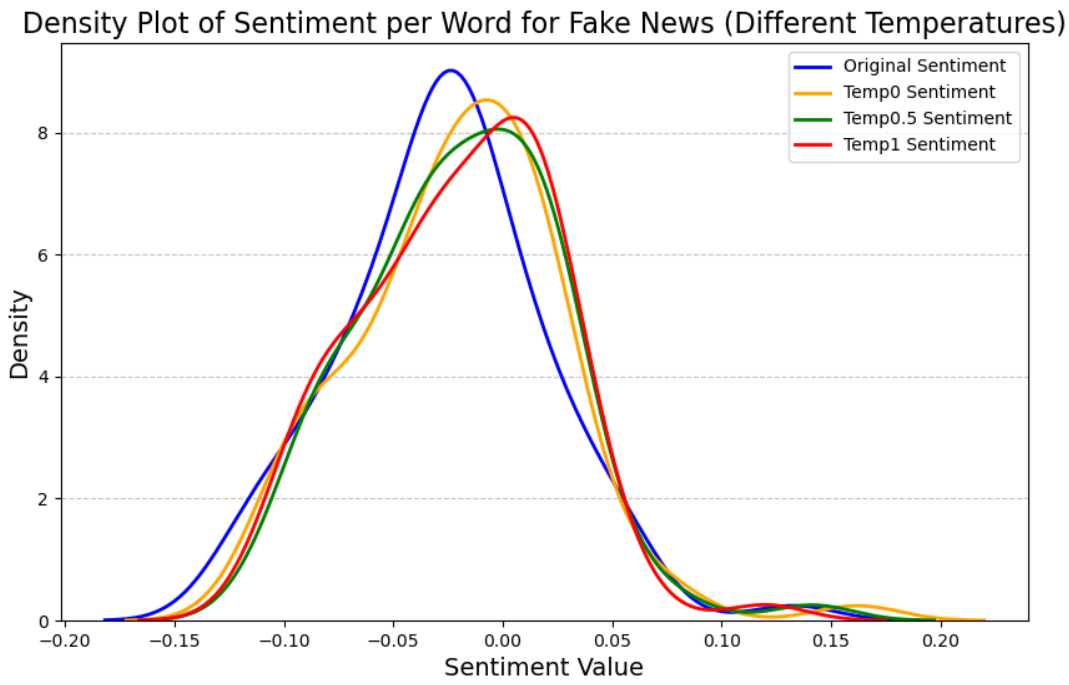

Figure 3: Graphique de densité du score sentimental par mot pour les fausses nouvelles, regroupée par température.

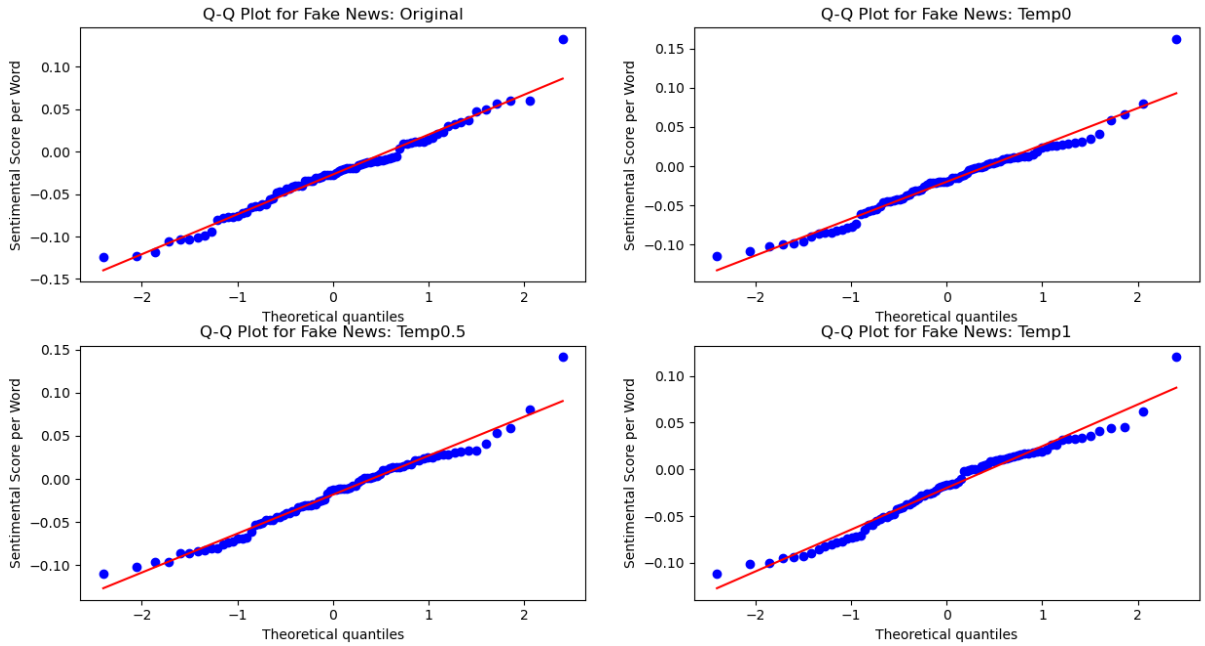

Figure 4: Tracé QQ du score sentimental par mot pour les fausses nouvelles, regroupées par température.

On a utilisé la bibliothèque SciPy pour effectuer le test de Shapiro-Wilk sur les données associées à chaque température, afin de vérifier leur normalité. Les hypothèses sont :

- H_0: *Les données suivent une distribution normale.*
- H_1: *Les données suivent une distribution non − normale.*

Avec un niveau de signification de 0.01 (1%), où on rejette H_0 si p < 0.01, on obtient les résultats suivants :

Tableau 4: Test de normalité Shapiro-Wilk sur le score sentimental par mot pour les fausses nouvelles.

Groupe de données : fausses nouvelles	Statistique de test W	Value de p	Résultat
Original	0.9813	0.25627	Accepter H_0
Température ChatGPT 0.0	0.9642	0.018614	Accepter H_0
Température ChatGPT 0.5	0.9752	0.10060	Accepter H_0
Température ChatGPT 1.0	0.9752	0.10077	Accepter H_0

Dans chaque cas, la valeur de p est plus grande que le niveau de signification de 0.01. Alors, on accepte H_0 et on assume que les données suivent une distribution normale.

Articles vrais

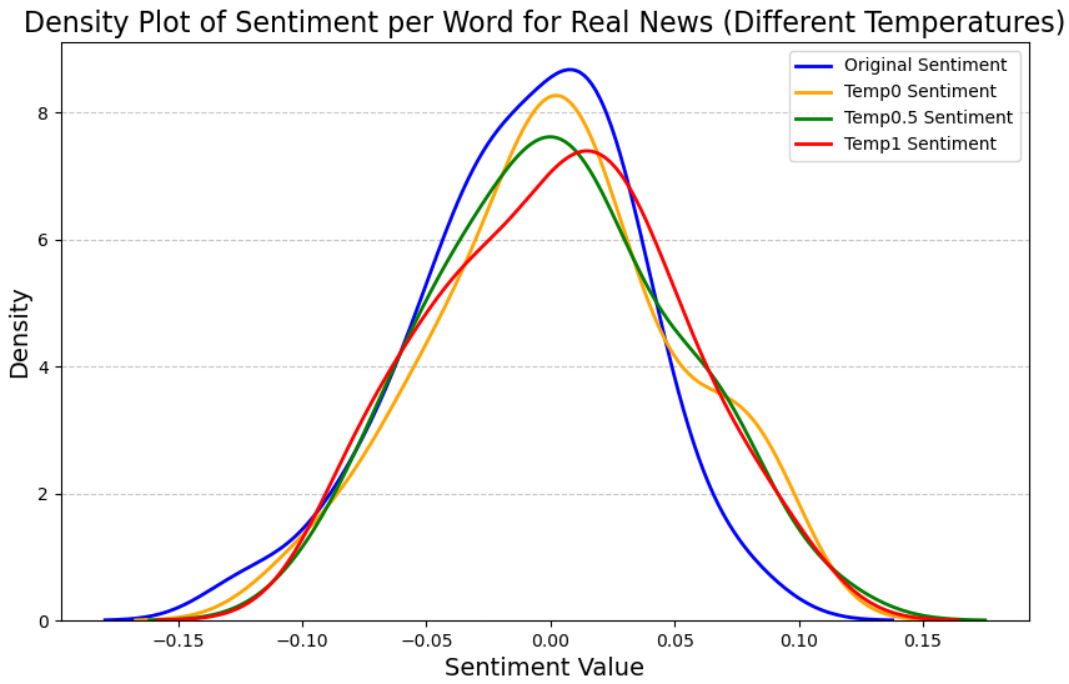

Figure 5: Graphique de densité du score sentimental par mot pour les vraies nouvelles, regroupée par température.

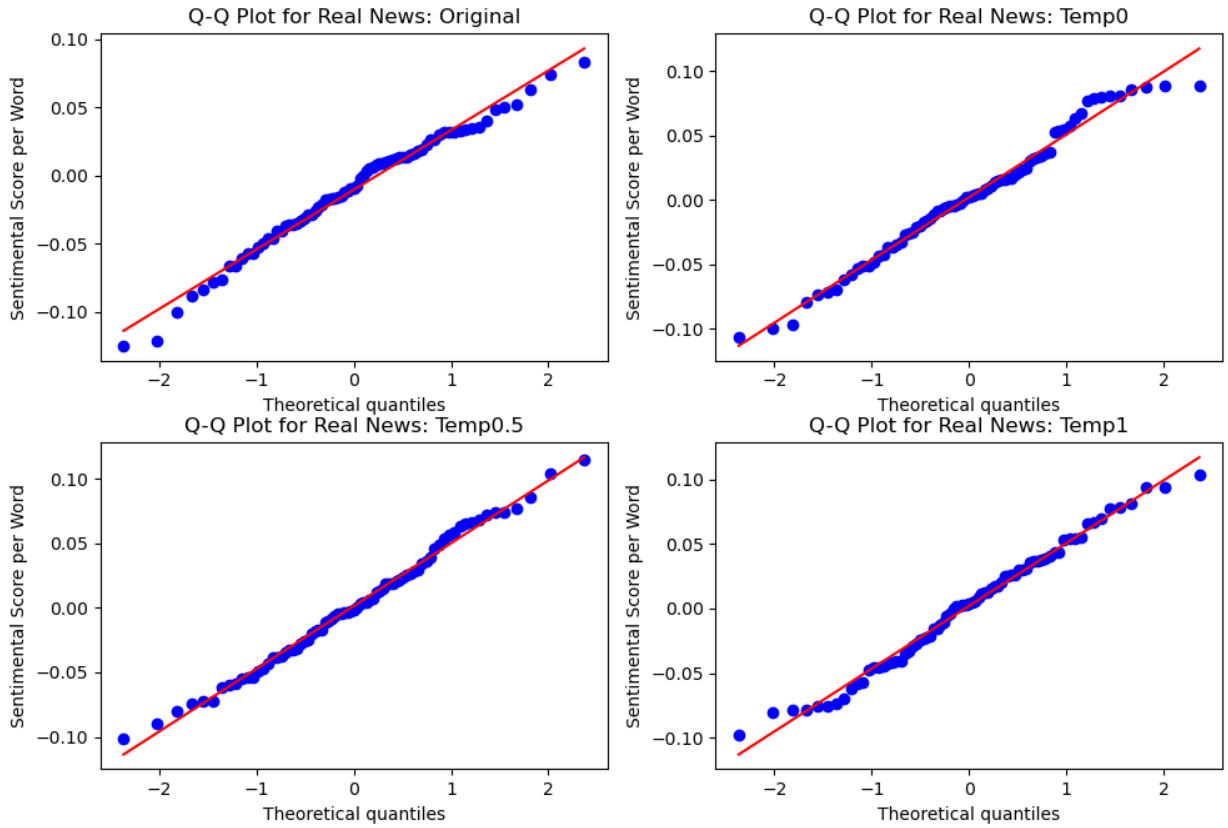

Figure 6: Tracé QQ du score sentimental par mot pour les vraies nouvelles, regroupées par température.

Tableau 5: Test de normalité Shapiro-Wilk sur le score sentimental par mot pour les vraies nouvelles.

Groupe de données : vraies nouvelles	Statistique de test W	Value de p	Résultat
Original	0.9830	0.39331	Accepter H_0
Température ChatGPT 0.0	0.9788	0.22584	Accepter H_0
Température ChatGPT 0.5	0.9904	0.83775	Accepter H_0
Température ChatGPT 1.0	0.9847	0.48219	Accepter H_0

On peut affirmer que dans les articles originaux et générés, le score sentimental par mot suit une distribution normale.

La température assignée à ChatGPT n'affecte pas l'objectivité du langage utilisé

Dans l'API de ChatGPT, il est possible d'assigner un paramètre de température à ChatGPT. Ce paramètre peut prendre des valeurs : [0, 1]. Ce paramètre influence la créativité et la diversité des réponses générées. En théorie, plus la température est élevée, plus ChatGPT est libre d'utiliser un langage expressif. On cherche à déterminer si la température affecte l'objectivité du langage utilisé lors de la génération d'articles.

On a calculé la différence entre le score sentimental par mot original et le score sentimental par mot généré de chaque article :

$$différence\ de\ score\ sentimental = \left|S_{orig} - S_{gen}\right|$$

Ensuite, on a fait un test ANOVA, avec trois listes de données : différence de score sentimental entre les articles originaux et les articles générés avec une température de 0.0, de 0.5 et de 1.0. Le test ANOVA a été fait en Python avec la bibliothèque SciPy.

- H_0: $Les\ différences\ de\ scores\ sentimentaux\ sont\ égales.$
- H_1: $Au\ moins\ une\ température\ a\ un\ différence\ de\ score\ sentimental\ différent.$

Avec un niveau de signification de 0.05, on peut affirmer que pour les fausses nouvelles tant que pour les vraies nouvelles, les différences de scores sentimentaux sont égales entre les températures.

Pour les fausses nouvelles, on a un F-score de 0.1407 et une valeur de p de 0.8689.
Pour les vraies nouvelles, on a un F-score de 0.0013 et une valeur de p de 0.9987.

On accepte H_0, puisque la valeur de p est supérieure à 0.05, ce qui indique qu'il n'y a pas de différence significative entre les températures.

Un F-score proche de 0 indique que la variance des données au sein de chaque groupe est plus importante que la variance des données entre les groupes. C'est-à-dire qu'il y a plus de variabilité dans les scores sentimentaux des articles générés avec une même température (par exemple, les différences de scores sentimentaux au sein des articles générés avec une température de 0.5) que dans les scores sentimentaux entre les différentes températures (c'est-à-dire les différences entre les scores sentimentaux générés à température 0.0, 0.5 et 1.0).

En d'autres termes, la variabilité au sein des articles générés par chaque température est beaucoup plus grande que la variabilité observée entre les températures elles-mêmes. Cela suggère qu'il n'y a pas de différence significative entre les moyennes des scores sentimentaux des différentes températures selon les résultats de l'ANOVA, et donc que les variations dans les scores sentimentaux sont plus dues à la variabilité interne des articles qu'à l'effet de la température utilisée dans la génération des textes.

ChatGPT génère des articles en utilisant un langage qui reste aligné avec l'objectivité des articles originaux

Si l'article original utilise un langage négatif, ChatGPT génère un nouvel article suivant cette tendance, bien que ChatGPT utilise souvent un langage légèrement plus positif. Il est improbable que ChatGPT, avec un article négatif comme référence, utilise un langage très positif, à l'exception de cas aberrants.

On a confirmé le comportement de ChatGPT, soit que ChatGPT suit le langage des articles originaux, en effectuant un test de Spearman avec la bibliothèque SciPy.

Le test de Spearman prend en entrée deux variables et retourne une valeur de corrélation de rang. Cette corrélation peut avoir une valeur entre -1 et 1. Avec un niveau de signification de 0.05, si l'indice de corrélation est plus grand que 0.05, alors il y a une relation monotone positive entre les deux variables. C'est-à-dire, si la première variable augmente, la deuxième augmente également.

- H_0: *Les scores sentimentaux originaux et générés sont indépendants.* $(\rho = 0)$
- H_1: *Les scores sentimentaux originaux et générés sont dépendants.* $(\rho \neq 0)$

Tableau 6: Corrélation de Spearman pour les scores sentimentaux par mot pour les fausses nouvelles.

Variables (scores sentimentaux par mot pour fausses nouvelles)	Corrélation de Spearman (ρ)	Résultat
Original et Température 0.0	0.7637	Dépendant
Original et Température 0.5	0.7872	Dépendant
Original et Température 1.0	0.7854	Dépendant

Tableau 7: Corrélation de Spearman pour les scores sentimentaux par mot pour les vraies nouvelles.

Variables (scores sentimentaux par mot pour vraies nouvelles)	Corrélation de Spearman (ρ)	Résultat
Original et Température 0.0	0.8669	Dépendant
Original et Température 0.5	0.8828	Dépendant
Original et Température 1.0	0.8806	Dépendant

On peut affirmer que le score sentimental généré par ChatGPT est proportionnel et dépendant du score sentimental original/introduit comme source. Cela signifie que si l'article original est négatif, l'article généré sera également négatif. Les articles générés par ChatGPT risquent de tendre légèrement vers l'usage d'un langage plus positif, tel qu'examiné plus tôt, sans être excessivement positif ou négatif.

Il n'est pas possible d'affirmer que ChatGPT génère des articles avec un biais politique

On cherche à déterminer si ChatGPT présente un biais politique, en particulier quant au score sentimental. L'objectif est d'évaluer si le langage généré par ChatGPT favorise un côté du spectre politique, gauche ou droite.

La base de données d'articles de FakeNewsNet présente les articles analysés par Buzzfeed, séparés en deux sections : fausses nouvelles et vraies nouvelles. Chacune de ces catégories est ensuite subdivisée en deux sous-catégories : sources ayant un biais politique de gauche et sources avec un biais politique de droite. Cette classification repose sur des évaluations externes de la crédibilité et du biais des nouvelles, notamment sur des plateformes telles que Media Bias Fact Check, lorsque cela est possible. En l'absence de classification provenant d'une plateforme externe, l'article a été soit omis, soit manuellement classifié après une brève lecture et analyse de l'article original.

Voici les résultats obtenus après la subdivision des données en gauche et droite :

Tableau 8: Moyennes des scores sentimentaux par mot pour les fausses nouvelles, séparées par biais

Fausses nouvelles	Original	Température ChatGPT 0.0	Température ChatGPT 0.5	Température ChatGPT 1.0	Généré
Score sentimental par mot (gauche, 14 articles)	-0.03371	-0.02576	-0.02281	-0.0214	-0.02332
Score sentimental par article (gauche, 14 articles)	-22.5	-16.5857	-16.9857	-16.2857	-16.6190
Score sentimental par mot (droite, 64 articles)	-0.02291	-0.01676	-0.01512	-0.01745	-0.01644
Score sentimental par article (droite, 64 articles)	-11.1719	-6.98125	-6.23125	-7.48125	-6.89792

Tableau 9: Moyennes des scores sentimentaux par mot pour les vraies nouvelles, séparées par biais

Vraies nouvelles	Original	Température ChatGPT 0.0	Température ChatGPT 0.5	Température ChatGPT 1.0	Généré
Score sentimental par mot (gauche, 9 articles)	0.005513	0.025214	0.020919	0.023716	0.023283
Score sentimental par article (gauche, 9 articles)	2.777778	11.11111	9.311111	10.15556	10.19259
Score sentimental par mot (droite, 14 articles)	-0.0243	-0.01773	-0.01349	-0.02288	-0.01803
Score sentimental par article (droite, 14 articles)	-14.4286	-8.81429	-7.27143	-12.2714	-9.4532

On cherche premièrement à déterminer s'il y a une différence d'objectivité différent avec les articles générés de gauche comparés avec les articles générés de droite. Les formules sont les suivantes :

$$Différence\ d'objectivité\ relatif\ (pourcentage) = \frac{(S_{gen} - S_{orig})}{|S_{orig}|}\ x\ 100$$

Où :

- S_{gen} = score sentimental généré
- S_{orig} = score sentimental original

Et :

$$Changement\ d'objectivité\ absolu = S_{gen} - S_{orig}$$

Tableau 10: Différence d'objectivité des articles générés par ChatGPT avec les articles originaux, séparés par biais, par mot

Type de nouvelles	Différence d'objectivité relatif (%)	Différence d'objectivité absolu
Fausses nouvelles, gauche (par mot)	38.16	0.01439
Fausses nouvelles, droite (par mot)	28.24	0.00647
Vraies nouvelles, gauche (par mot)	322.3	0.01777
Vraies nouvelles, droite (par mot)	25.80	0.00627

Tableau 11: Différence d'objectivité des articles générés par ChatGPT avec les articles originaux, séparés par biais, par article

Type de nouvelles	Différence d'objectivité relatif (%)	Différence d'objectivité absolu
Fausses nouvelles, gauche (par article)	25.81	5.809
Fausses nouvelles, droite (par article)	38.25	4.2740
Vraies nouvelles, gauche (par article)	266.9	7.4148
Vraies nouvelles, droite (par article)	34.48	4.9754

Cette expérience comporte des limitations et peut être améliorée. Idéalement, on aurait davantage d'articles à analyser, en particulier provenant de sources projetant un biais politique de la gauche, avec des articles de la gauche et de la droite de taille uniforme et des scores sentimentaux originaux comparables. Cela étant dit, la tendance observée chez ChatGPT se présente dans la mesure qu'il utiliserait un langage plus positif pour la gauche et ce, autant pour les vraies nouvelles que pour les fausses nouvelles.

En termes de différence d'objectivité absolu, pour les fausses nouvelles, chaque mot généré par ChatGPT est environ 2.2 fois plus positif pour la gauche que pour la droite, tandis que pour les vraies nouvelles, ce ratio monte à environ 2.8 fois plus positif pour la gauche.

$$augmentation\ relatif = \frac{changement\ d'objectivité\ absolu\ par\ mot\ (gauche)}{changement\ d'objectivité\ absolu\ par\ mot\ (droite)}$$

Ce phénomène peut être visualisé avec l'aide des boîtes à moustaches :

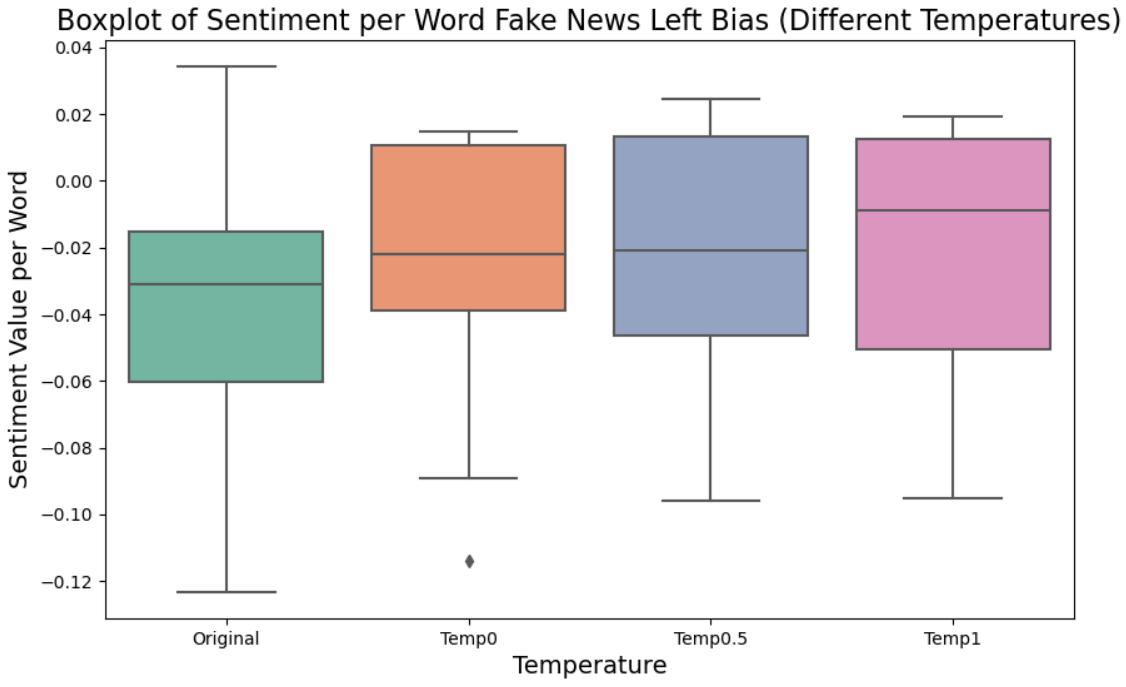

Figure 7: Boîte à moustaches du score sentimentale par mot pour les fausses nouvelles, biais de gauche, regroupée par température.

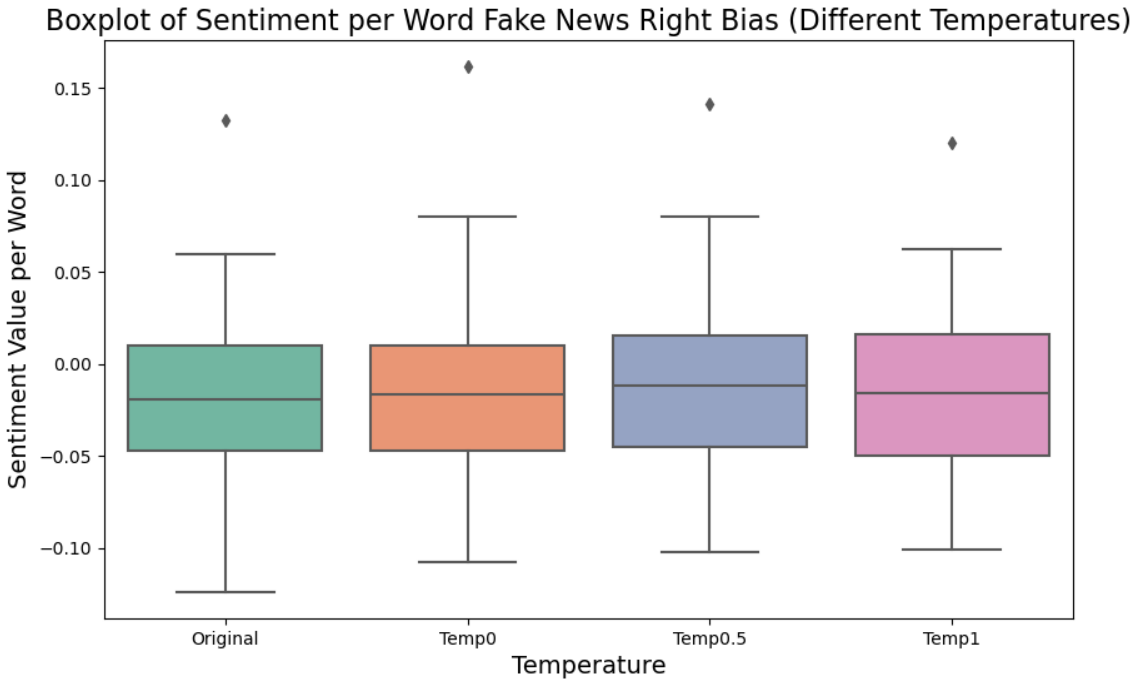

Figure 8: Boîte à moustaches du score sentimentale par mot pour les fausses nouvelles, biais de droite, regroupée par température.

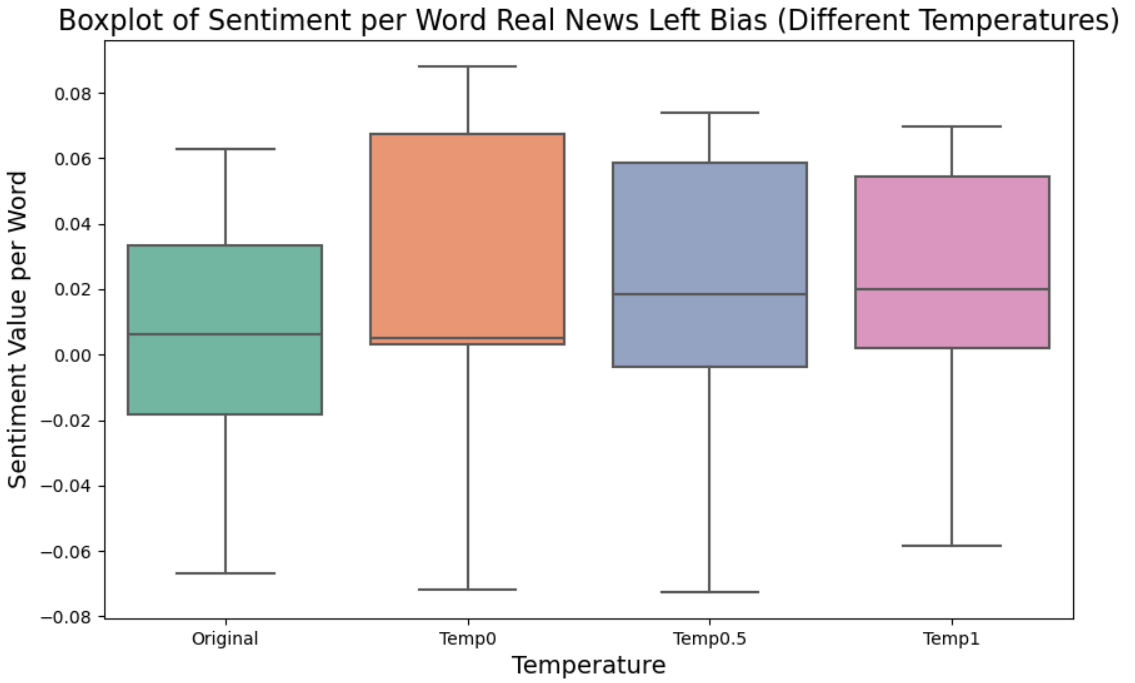

Figure 9: Boîte à moustaches du score sentimentale par mot pour les vraies nouvelles, biais de gauche, regroupée par température.

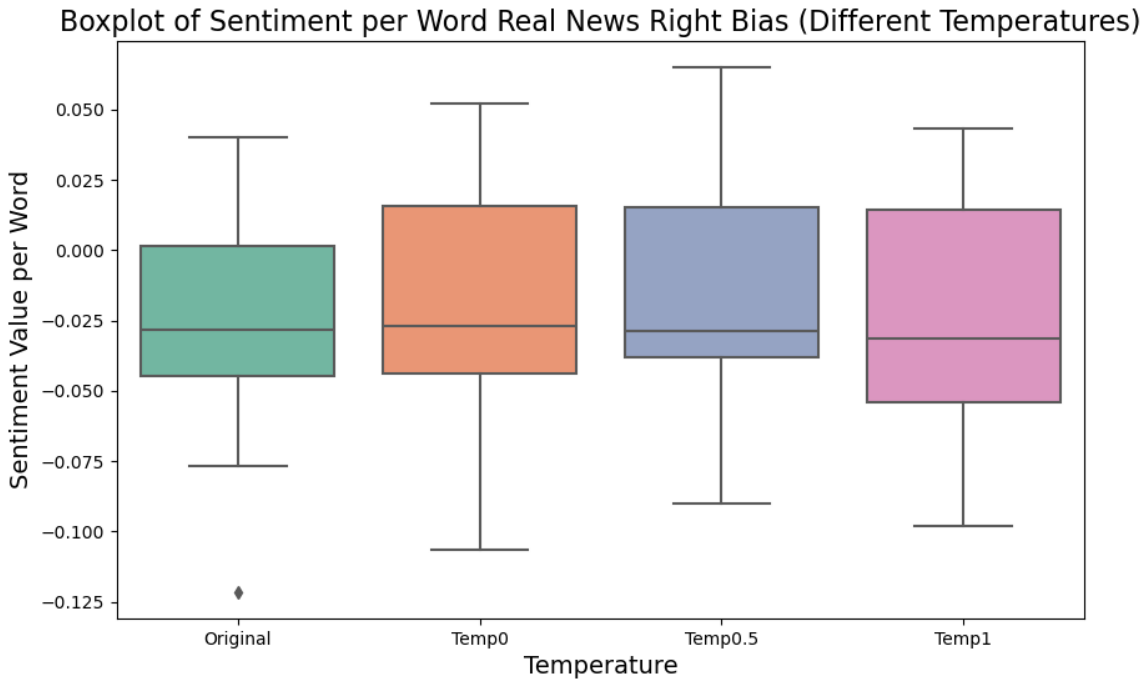

Figure 10: Boîte à moustaches du score sentimentale par mot pour les vraies nouvelles, biais de droite, regroupée par température.

Les résultats suggèrent que ChatGPT pourrait avoir tendance à utiliser un langage plus positif dans certains articles associés à des sources biaisées à gauche. Cependant, l'échantillon est restreint et souvent les articles sont négatifs envers la droite. Ainsi, puisque ChatGPT utilise un langage plus positif, les résultats montrent que ChatGPT utilise un langage perçu comme plus modéré dans certains cas.

Études de cas

En utilisant la méthode IQR, on a identifié 4 articles générés de fausses nouvelles et 4 articles générés de vraies nouvelles où le score sentimental par mot généré diffère considérablement de celui des articles originaux. En utilisant la méthode du score Z, on a identifié 2 articles générés de fausses nouvelles et 2 articles générés de vraies nouvelles où le score sentimental par mot généré diffère considérablement avec celui des articles originaux.

La méthode IQR et du score Z ont trouvé deux articles générés de fausses nouvelles et deux articles générés de vraies nouvelles en commun qui diffèrent considérablement avec celui des articles originaux. On remarque que trois d'entre eux proviennent de sources de la droite et un article provient d'une source de la gauche/neutre.

Tableau 12: Articles générés avec un score sentimental par mot d'une différence extrême comparé à celui des articles originaux

ID	Original (article/per word)	Temp 0 (article/per word)	Temp 0.5 (article/per word)	Temp 1 (article/per word)
Fake_27-Webpage	-52/-0.07669	7.2/0.009595	15.8/0.02132	6.4/0.009248
Fake_51-Webpage	-5/-0.01126	-26.8/-0.06077	-35.8/-0.0805	-46/-0.1004
Real_62-Webpage	10/0.01852	48/0.08111	61.8/0.1039	56.8/0.09408
Real_9-Webpage	6/0.02956	18.6/0.08589	25.2/0.1148	22.8/0.1039

Voici un extrait du faux article 27 original intitulé : « How Democrats are Going to Try to STEAL the Election for Hillary » et ce que ChatGPT a généré.

Original :
« Due to the unfortunate fact that the Democratic Party, the Obama administration, our treasonous Congress, and our abominably corrupt federal court system, have suppressed any and all attempts to enact laws and procedures to ensure fair and honest elections, the American people have as our only recourse in the November election, to volunteer en mass to monitor the polling places across our nation to prevent voter fraud to the greatest extent possible. »

Tableau 13: Information sur le faux article original 27

Nombre de mots	Score sentimental	Score sentimental par mot
79	-3.0	-0.03797

Mots avec un score important :

- « treasonous », -3.0
- « corrupt », -3.0
- « fair », 2.0
- « honest », 2.0
- « fraud », -4.0
- « greatest », 3.0

Généré par ChatGPT température 0.5, itération 1 :
« The author further alleges that various institutions, including the Democratic Party, the Obama administration, Congress, and federal courts, are actively suppressing fair election practices. This claim underscores a growing concern among certain segments of the population regarding the perceived erosion of electoral integrity. It is within this context that the author calls upon citizens to take an active role in safeguarding the democratic process. The author encourages individuals to volunteer as monitors at polling places during the upcoming elections, emphasizing the importance of vigilance in preventing voter fraud. »

Tableau 14: Information sur le faux article généré 27, température 0.5, itération 1

Nombre de mots	Score sentimental	Score sentimental par mot
98	6.0	0.06122

Mots avec un score important :

- « fair », 2.0
- « growing », 1.0
- « integrity », 2.0
- « active », 1.0
- « importance », 2.0
- « fraud », -4.0

Une théorie possible expliquant pourquoi l'article généré utilise un langage plus positif est que ChatGPT a choisi d'écrire à la troisième personne. Cette approche n'est pas courante, car, généralement, ChatGPT génère des textes à la première personne. Il serait intéressant d'étudier plus profondément le comportement de ChatGPT concernant la génération d'articles du point de vue de la première personne comparé à un point de vue de la troisième personne.

Voici un extrait du faux article 51 original intitulé : « Two White Men Doused With Gasoline, Set On FIRE By Blacks » et ce que ChatGPT a généré.

Original :
« What goes through someone's deplorable brain when they're thinking "well right about now is a great time to light these guys on fire. Yup, let's just toss this cocktail over on them and watch the men burn."

Who carries the items needed for a good Molotov cocktail anyway? Was this premeditated? It's not like I have the items needed for a good cocktail laying around unless it's a Bloody Mary on a Sunday morning. »

Tableau 15: Information sur le faux article original 51

Nombre de mots	Score sentimental	Score sentimental par mot
91	6.0	0.06593

Mots avec un score important :

- « great », 3.0
- « fire », -2.0
- « good », 3.0
- « like », 2.0
- « Bloody », -3.0

Généré par ChatGPT température 1, itération 3 :

« The author reflects deeply on the violent act, specifically questioning the mindset that prompts individuals to resort to extreme measures, including the use of a Molotov cocktail. The act of using such a destructive tool not only signifies a departure from rationality but also evokes a broader discussion about the normalization of violence in society today. »

Tableau 16: Information sur le faux article généré 51, température 1, itération 3

Nombre de mots	Score sentimental	Score sentimental par mot
60	-10.0	-0.1667

Mots avec un score important :

- « violent », -3.0
- « questioning », -1.0
- « destructive », -3.0
- « violence », -3.0

Il semble que l'algorithme Afinn rencontre des difficultés au niveau de la compréhension du sarcasme. Par exemple, l'expression « good Molotov cocktail » obtient un score sentimental de +3.0, puisqu'elle contient le mot « good ». De la même manière, bien que ChatGPT comprenne le sarcasme et interprète le message de l'article original, ChatGPT ne reproduit pas le sarcasme. En effet, ChatGPT décrit la situation de manière directe, en utilisant des mots négatifs, tandis que l'article original a recours intentionnellement au sarcasme pour communiquer son message d'une façon particulière.

Le vrai article 62 contient principalement des citations d'une personne qui soutient Trump. Les articles générés par ChatGPT expliquent le sentiment des citations positivement.

Le dernier article, soit l'article vrai 9, est intéressant car il provient de CNN, une source du centre/de la gauche. L'article généré par ChatGPT met en valeur les adjectifs ajoutés par ChatGPT, donnant une connotation positive à l'article.

Original :

« Clinton's strength comes from the Atlanta area, where she leads Trump 55% to 35%. But Trump leads her 51% to 33% elsewhere in the Peach State. She leads 88% to 4% among the state's black voters, but trails 20% to 66% among white voters. »

Tableau 17: Information sur le vrai article original 9

Nombre de mots	Score sentimental	Score sentimental par mot
59	2.0	0.03390

Mots avec un score important :

- « strength », 2.0

Généré par ChatGPT température 0.5, itération 3 :

« Clinton's strongest support is concentrated in the Atlanta area, where she leads Trump significantly, 55% to 35%. However, outside of Atlanta, Trump holds a commanding lead, winning 51% to 33%. The demographic breakdown shows that Clinton has overwhelming backing from black voters, with 88% support, while Trump receives 66% of the white voter demographic. »

Tableau 18: : Information sur le vrai article généré 9, température 0.5, itération 3

Nombre de mots	Score sentimental	Score sentimental par mot
71	12.0	0.1690

Mots avec un score important :

- « strongest », 2.0
- « support », 2.0
- « winning », 4.0
- « backing », 2.0

On observe que ChatGPT ajoute beaucoup d'adjectifs positifs. ChatGPT ajoute également des mots qui ne sont pas jugés positifs par Afinn, car l'algorithme manque de contexte, tels que « leads Trump significantly », « commanding lead », « overwhelming ».

Ces articles aberrants montrent que l'objectivité du langage utilisé par ChatGPT pourrait dépendre de la conjugaison des verbes, plus précisément de la personne employée, comme la première ou la troisième personne. Ce comportement serait intéressant à analyser davantage dans le futur.

On remarque également que l'algorithme d'analyse sentimental Afinn rencontre des difficultés pour calculer le score sentimental lorsque le contexte est complexe ou que le texte comporte du sarcasme. ChatGPT est capable de comprendre le sarcasme, mais il semble avoir des difficultés à le reproduire sans une requête spécifique à cet effet. ChatGPT ajoute souvent des adjectifs qui introduisent une nuance positive dans les textes, ce qui pourrait apporter un biais.

ChatGPT écrit plus de mots, le plus haut la température

Le nombre d'articles analysés a été réduit en raison de contraintes liées à la longueur des textes générés et à la complexité de l'analyse. Lors de la génération des articles, on a spécifié à ChatGPT d'écrire environ le même nombre de mots que l'article original. Cependant, on a observé que ChatGPT rencontre une grande difficulté à générer au-delà de 1000 mots, quelle que soit la température, même lorsque cela est explicitement demandé. Néanmoins, on constate une relation monotone positive entre la température et le nombre de mots générés.

Tableau 19: Nombre de mots par article pour les fausses nouvelles, regroupés par température (85 articles)

Nombre de mots par article	Original	Température ChatGPT 0.0	Température ChatGPT 0.5	Température ChatGPT 1.0
Moyenne	509	503.4918	512.0494	522.3176
Médiane	476	495.6	511.2	509.6
Minimum	175	173.6	173.6	176
Maximum	1165	892.2	1089.6	1041.2

Tableau 20: Nombre de mots par article pour les vraies nouvelles, regroupés par température (77 articles)

Nombre de mots par article	Original	Température ChatGPT 0.0	Température ChatGPT 0.5	Température ChatGPT 1.0
Moyenne	457.5195	447.7532	449.5662	460.6831
Médiane	419	428.6	437	436
Minimum	140	135.2	139.4	141.2
Maximum	1178	978.2	1003.4	992.4

Dans les deux cas, on observe une augmentation du nombre de mots avec chaque augmentation de température. Il est important de noter que le nombre de mots utilisé pour chaque article (85 ou 77 articles) représente déjà la moyenne des résultats de 5 itérations de génération d'articles.

En d'autres termes, pour chaque article et pour chaque température, on a calculé la moyenne du nombre de mots générés.

$$Nb\ de\ mots\ généré = \frac{\sum_{i=1}^{5} Nb\ de\ mots\ dans\ l'article\ généré\ durnat\ l'itération\ i}{5}$$

Ensuite, on a calculé la moyenne du nombre de mots pour l'ensemble des articles à chaque température.

$$Moyenne\ du\ nombre\ de\ mots\ pour\ température\ x = \frac{\sum_{a=1}^{N} Nb\ de\ mots\ de\ l'article\ généré\ a}{N}$$

Où :

- $x = \{0.0, 0.5, 1.0\}$
- N = nombre d'articles traités (85 pour les fausses nouvelles et 77 pour les vraies nouvelles)

Discussion

Cette recherche a permis de visualiser le comportement de ChatGPT concernant la génération de contenu, et dans ce cas, d'articles de nouvelles en se basant sur des instructions et des données en format de puces. L'hypothèse de départ, que ChatGPT génère du contenu avec un biais politique favorisant la gauche ne peut pas être confirmée. Certes, on remarque que ChatGPT utilise un langage plus positif avec les articles de la gauche, mais ceci ne signifie pas nécessairement que ChatGPT favorise la gauche. Cela pourrait également refléter le fait que les articles originaux utilisent un langage particulièrement négatif pour décrire la droite et les politiciens tel Trump, par exemple, et que ChatGPT utilise un langage plus modéré pour le décrire. Cependant, il y a un manque d'articles pour parvenir concrètement à une conclusion. Les paramètres de départ ne sont pas idéaux : les scores sentimentaux des articles originaux de gauche et de droite ne sont pas égaux et la taille des articles originaux diffère grandement.

Il est intéressant de noter que le score sentimental des articles suit une distribution normale. Cela pourrait indiquer que les événements ou les thèmes couverts dans les nouvelles ont tendance à produire une diversité équilibrée d'émotions, allant du négatif au positif, avec une majorité de contenus restant neutres, ou légèrement négatives dans le cas des articles examinés durant cette recherche.

Il est important de remarquer que ChatGPT utilise un langage plus positif. Cela peut être dû à plusieurs raisons, comme le fait que ChatGPT tend à ajouter des adjectifs positifs pour décrire certaines statistiques ou peut-être qu'il existe une différence de choix de mots lorsque ChatGPT écrit à la première ou à la troisième personne.

ChatGPT ne reproduit pas le sarcasme des articles originaux, mais tend plutôt à expliciter le sentiment sous-jacent. Si l'article original emploie le sarcasme, comme vu dans l'étude de cas, ChatGPT peut employer un langage considéré plus négatif par les algorithmes d'analyse sentimentale, car le sarcasme utilise fréquemment des mots positifs dans un contexte négatif, quelque chose qu'au moins l'algorithme qu'on a utilisé, Afinn, ne peut pas identifier.

En général, selon le test de Spearman, ChatGPT génère des articles dont l'objectivité suit la tendance dans les articles originaux, c'est-à-dire que si un article original utilise un langage positif, l'article généré par ChatGPT utilisera également un langage positif.
On peut affirmer, à la suite d'un test d'ANOVA, que la température (degré de liberté d'expression) de ChatGPT a un effet négligeable concernant le score sentimental des articles générés. Cependant, on peut observer que ChatGPT a tendance à générer plus de mots lorsque la température augmente. De plus, ChatGPT a de la difficulté à générer plus de 1000 mots dans une requête concernant la génération d'articles.

Une autre amélioration à apporter à cette recherche serait la création d'un calcul de score sentimental qui donne plus de poids à la quantité de mots et moins à chaque article : les articles avec moins de mots valent moins, ainsi on peut réduire l'influence des cas extrêmes.

Cette étude pousse la recherche sur ChatGPT, l'objectivité de ChatGPT et le biais de ChatGPT en utilisant une nouvelle méthode. Cette méthode qui consiste à générer des articles en se basant sur des articles originaux, réels, avec l'API de ChatGPT, puis à analyser le score sentimental de ces articles avec Afinn, suivi d'une interprétation des résultats en utilisant des méthodes statistiques. Il serait intéressant de poursuivre cette recherche en se concentrant sur les différents thèmes et questions abordés et soulevés, entre autres, l'influence de la température de ChatGPT concernant la taille du contenu généré et le biais de ChatGPT avec des données plus uniformes.

Je vous invite de parcourir les données et les articles générés vous-même. Les données ainsi que le code sont disponibles sur github : https://github.com/hechuno/chatGPT-sentiment-analysis. Vous pouvez me contacter à l'adresse suivante pour toute question ou commentaire.
hechunou@gmail.com

Bibliographie

[1] Abid, Abubakar, Maheen Farooqi, et James Zou. « Persistent Anti-Muslim Bias in Large Language Models ». Dans Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021.

[2] Davis, J., Van Bulck, L., Durieux, B. N., & Lindvall, C. (2024). The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research. *JMIR human factors*, *11*, e53559. https://doi.org/10.2196/53559

[3] Depak, M. « FakeNewsNet ». Consulté le 22 octobre 2024. https://www.kaggle.com/datasets/mdepak/fakenewsnet?resource=download.

[4] Emir, S. « Israel Hamas Conflict News Dataset ». Consulté le 22 octobre 2024. https://www.kaggle.com/datasets/emirslspr/israel-hamas-conflict-news-dataset.

[5] Esuli, Andrea, et Fabrizio Sebastiani. « SentiWordNet ». Consulté le 20 septembre 2024. https://github.com/aesuli/SentiWordNet.

[6] Mohammad, Saif. « NRC Emotion Lexicon ». Consulté le 20 septembre 2024. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

[7] Nadeem, Moin, Anna Bethke, et Siva Reddy. « StereoSet: Measuring stereotypical bias in pretrained language models ». arXiv preprint arXiv:2004.09456 (2020).

[8] Nielsen, Finn Årup. « A new ANEW: évaluation d'une liste de mots pour l'analyse des sentiments dans les microblogs ». Dans Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages, vol. 718 de CEUR Workshop Proceedings, pp. 93–98, mai 2011. Édité par Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, et Mariann Hardey.

[9] Peutz, Steven. « Misinformation Fake News Text Dataset 79k ». Consulté le 22 octobre 2024. https://www.kaggle.com/datasets/stevenpeutz/misinformation-fake-news-text-dataset-79k.

[10] Recasens, Marta, Cristian Danescu-Niculescu-Mizil, et Dan Jurafsky. « Linguistic Models for Analyzing and Detecting Biased Language ». Dans Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013.

[11] SciPy Community. (n.d.). *SciPy documentation*. SciPy. https://docs.scipy.org/doc/scipy/

[12] Wilson, Theresa, Janyce Wiebe, et Paul Hoffmann. « MPQA Subjectivity Lexicon ». Multi-Perspective Question Answering Lexicons. Université de Pittsburgh. Consulté le 20 septembre 2024. http://mpqa.cs.pitt.edu/lexicons/.