

Improving Detection of ChatGPT-Generated Fake Science Using Real Publication Text: Introducing xFakeBibs a Supervised Learning Network Algorithm

Ahmed Abdeen Hamed

Sano Centre from Computational Medicine, Cracow, Poland

Correspondence: a.hamed@sanoscience.org

Abstract: Background: ChatGPT is becoming a new reality. Where do we go from here? **Objective:** to show how we can distinguish ChatGPT-generated publications from counterparts produced by scientist. **Methods:** By means of a newly devised algorithm, called xFakeBibs, we show the difference in contents and structure of bigram networks generated from ChatGPT fake publications are significantly different from real publications. Specifically, we triggered ChatGPT to generate 100 publications related to Alzheimer's and comorbidity. Using TF-IDF, we constructed a network of bigrams and compared with 10 other networks constructed from real PubMed publications. Each of those networks were constructed from exactly one of 10 folds. Each fold was equally comprised of 100 publications to ensure fairness. We trained the xFakeBibs algorithm using the 10-folds which were used to test the ChatGPT fake publications. The algorithm successfully assigned the POSITIVE label for real and NEGATIVE for fake ones. **Results:** When comparing the bigrams of the training set against all the other 10 folds, we found that the similarities fluctuated between (19%-21%). On the other hand, the bigram similarity from the ChatGPT was only (8%). Additionally, when testing how the bigrams alter the structure of the training model, we found that all 10 folds contributed (51%-70%) new, while ChatGPT contributed only 23% which is less than half of any of the other 10 folds. Upon calibrating the xFakeBibs algorithm using the 10-fold real publications, we learned that they contribute 21.96-24.93 number of edges on average. When xFakeBibs classified the individual articles, we found that 98 of the 100 publications were detected as fake, while 2 articles failed the test and were classified as real publications. **Conclusions:** While it is indeed possible to distinguish, in bulk, a dataset of ChatGPT-generated publications from counterparts of real publications (as was the case for the Alzheimer's dataset). Classifying individual articles as fake though can be done to a high degree of accuracy, it may not be possible to detect all fake and ChatGPT automatically generated articles. ChatGPT may seem to be a useful tool, but it certainly presents a threat to our knowledge of authentic and real science. This work is indeed a step in the right direction to counter fake science and misinformation.

Keywords: ChatGPT; Generative AI; Fake Publications; Human-Generated Publications; Supervised Learning; ML Algorithm; Fake Science; NeoNet Algorithm

1. Introduction

With ChatGPT being a new reality, our world is in a controversial state. On the one hand, there are those who see potential and seek to utilize it. On the other hand, there are those who remain skeptical and search for validation and further assessments to decide how this tool will affect the different aspects of our lives. This split provided a strong motivation for this work and triggered the effort of providing a tool that provides assessment to a significant resource of our lives which is scientific publications. Without any doubt, this resource is one of the most sacred sources of knowledge. That is because it is an invaluable source of any future discovery¹⁻³. The spread of various predatory journals particularly contributed to fake science and caused to be on the rise⁴⁻⁷. With the influence of social media, the impact is far-reaching^{8,9}. Particularly, during the Coronavirus global pandemic, the propagation of misinformation that surrounded the significance of vaccination had lead people to reject it and putting others at harm¹⁰⁻¹². Another disturbing example that occurred also during the pandemic was the propagation of a fake article that reported that the deficiency of

vitamin D led to 99 per cent of death of the studied population. Though the article was eventually withdrawn, the damage of the misinformation was magnified by a DailyMail news article and made it global^{13,14}. It is crucial to protect authenticity of the scientific work recoded in publications and others forms from fraud or any influential factors that make it an untrusted source of knowledge.

Here, we demonstrate how the emergence of ChatGPT (and many other Generative AI tools) have, in many ways, impacted our society today: (1) the launch of many special issues and themes to study, assess, analyze, and test the impact and potential of ChatGPT^{15–18}, (2) the adoption of new policies by journals regarding ChatGPT authorship^{19–23}, (3) the development of ChatGPT plugins, and inclusion in professional services such as Expedia and Slack²⁴, (4) the development of educational tools as in Wolfram and potential of creating support tool for learning as in the case for the Medical Licensing Examination²⁵.

2. Methods

2.1. Data Collection

We used two different datasets: (1) to train and calibrate the algorithm. By means of querying the PubMed portal using the query “Alzheimer’s and comorbidities”, we collected 1000 publications that are human-produced and archived in the digital repository (2) to test ChatGPT articles, we also asked ChatGPT to generate 100 articles with titles and abstracts that are also related to “Alzheimer’s and comorbidities”. Both datasets were preprocessed using the same mechanism for noise and stop words removal.

2.2. Statistical Analysis

We performed two different types of analysis to discriminate contents from real publications and contents generated from ChatGPT: one that analyzed bigrams of equal number of publications datasets, and the other comparing the structure of networks resulting from the bigrams generated.

2.3. ChatGPT Analysis

Using the Term Frequency-Inverse Term Frequency (TF-IDF) measure, we performed against three different types of sources (1) the training dataset, which is a slide of a 100 real publications, (2) the calibration using 10-folds each of which is also 100 publications, and (3) the testing of 100 publications generated using the ChatGPT. We compared the bigrams of the training dataset with all the 10-fold to calibrate before we tested against the ChatGPT data. We have computed the percentage of the bigram overlapping with the training and computed the similarities. We also calculated the ratio of the overlapping bigram against the 10-fold and the ChatGPT as another measure of comparison.

2.4. ChatGPT-Network Structural Analysis

Bigrams are any two consecutive words that may prove significant in any given text-based dataset^{26,27}. They also provide a significant mechanism for constructing networks of words^{14,28}. The purpose of this type of analysis was to study whether the bigram networks generated from real publication are significantly different in structure from the ChatGPT network counterpart. Here, we expose a specific structural property that has proven to discriminate a bigram network from real publication vs a counter part of fake publications generated from ChatGPT, namely computing the largest connected components (words of bigrams). Following the same approach above, we computed the largest connected component of the training dataset to establish a baseline of comparison, followed by computing the largest components of all the 10-folds to establish a calibration, before we can test against the bigram network of the ChatGPT. The following step describe the Algorithm 1 that we used for the calibration step.

1. Starting with the 10-folds, we computed the largest connected component, 1-fold at a time
2. We added the largest component of each fold to the training components one edge at a time (represented by the bigram)
3. We only considered the bigrams that altered the structure (i.e., the bigrams that introduced new edges to the original node of the training network) and ignored the bigrams that added new nodes.
4. We measured how the bigrams of each fold altered the training network, by calculated the following ratio:

$$\frac{\text{Number of Newly Introduced Edges} - \text{Number of Original Edges}}{\text{Number of Original Edges}}$$

5. The ratio offered a lower and upper bound to discriminate how the introduction of a guanine publication alters the training network
6. We measured the same ratio for the network generated from the ChatGPT and compared to all ratios generated from all 10-folds

Table 1. show the summary statistics generated from comparing the number of bigrams generated from 10-folds with the ChatGPT bigrams. The Data Source column labels shows specifies whether it is one of the 10-fold of the ChatGPT data. The next column captures the number of overlapping bigrams in the training when compared them with all other sources. The next column measures the percentage of the overlap compared to its own size. The last column summarizes the overlap as a percentage and presents it as a similarity between all the 10-fold on one hand and the ChatGPT on the other hand. While the number of overlapping bigrams in all the 10-folds fluctuated between (178-202) number of bigrams, the ChatGPT offered 81 bigrams overlap. This was summarized in a similarity percentage as (19%-22%) against ChatGPT which only shared 9% similarity.

Table 1. summarizes the statistics resulted from comparing the bigrams of training data, calibrating folds, and finally the ones generated from ChatGPT. The first column captures the data source where the bigrams generated, the second column displays the number of overlapping bigrams when compared with training, the third column shows the percentage of overlapping bigrams to self, while the last the similarity percentage to the training dataset.

	Data Source	Number of Bigrams Overlaps	Percent to self %	Similarity to Training%
0	Fold-1	202	0.21	0.22
1	Fold-2	183	0.19	0.20
2	Fold-3	178	0.22	0.19
3	Fold-4	178	0.20	0.19
4	Fold-5	180	0.21	0.20
5	Fold-6	180	0.19	0.20
6	Fold-7	179	0.19	0.19
7	Fold-8	181	0.20	0.20
8	Fold-9	184	0.20	0.20
9	Fold-10	201	0.23	0.22

10	GPT-Test	81	0.16	0.09
----	----------	----	------	------

We also summarize the statistics summarized in the Table 1 using a barplot in Figure 1. The X axis shows the data source (10-Folds and ChatGPT) while the Y axis shows the actual the number of nodes contributed scored by each source. The diagram shows a significant difference between 10-fold of real publications vs fake publications generated from the ChatGPT.

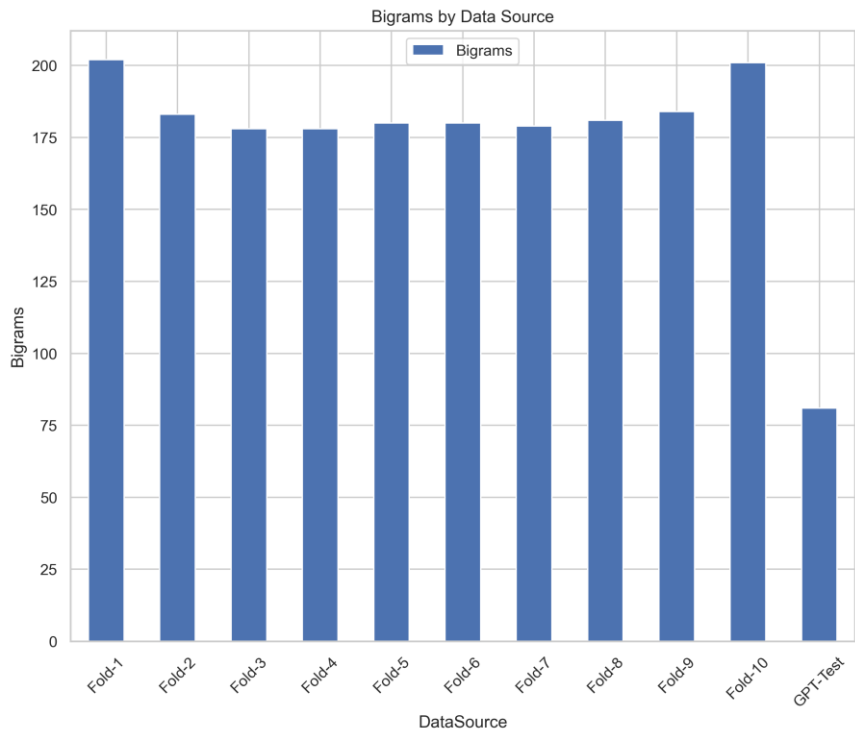


Figure 1. shows the number of overlapping bigrams of 10-folds of publications vs ChatGPT generated publications, when compared with a training dataset from real publications. .

3.2. Outcome of ChatGPT Structural Analysis: Largest Connected Components

The analysis described in the Methods section can also be summarized using Table 2 which is structured in columns as follows: Data Source, Number of Nodes, Number of Edges, Number of Connected Components, and Connected Component Percentage. The table indicates that the connected component percentages generated from the 10-folds of real publications lie between (51%-70%), ChatGPT scored only 23%. Clearly, there is a serious deficit of the ChatGPT contribution to the structure of the training data by at least approximately 50%.

Table 1 Shows the structural analysis of the ChatGPT network when compared with 10-folds bigram networks. In particular, the connected components percentage scored 23% when compared with all other 10-fold which scored (51%-70%).

<i>Index</i>	Data Source	No. Nodes	No. Edges	No. Connnected Components	Connected Components Percent%
<i>0</i>	F01	1086	1624	855	0.67
<i>1</i>	F02	1078	1625	867	0.69
<i>2</i>	F03	1019	1494	777	0.51
<i>3</i>	F04	1058	1570	827	0.61
<i>4</i>	F05	1068	1533	825	0.61
<i>5</i>	F06	1102	1624	860	0.68
<i>6</i>	F07	1077	1593	871	0.70
<i>7</i>	F08	1108	1580	846	0.65
<i>8</i>	F08	1108	1580	846	0.65
<i>9</i>	F10	1075	1540	848	0.65
<i>10</i>	ChatGPT	801	1312	632	0.23

Figure 2 also summarizes the full table statistics represented by the number of nodes, the number of edges, and the size of connected components. The figure shows an interesting behavior of how the size of the connected components were very closely clustered together, the size of ChatGPT connected components (represented by a blue dot at the bottom left of each plot) looked isolated and appears to be an insignificant anomaly. While Figure 3 focuses only the percentage of the connected components and how they ChatGPT contributes a much smaller slice when it is compared with all other 10-folds.

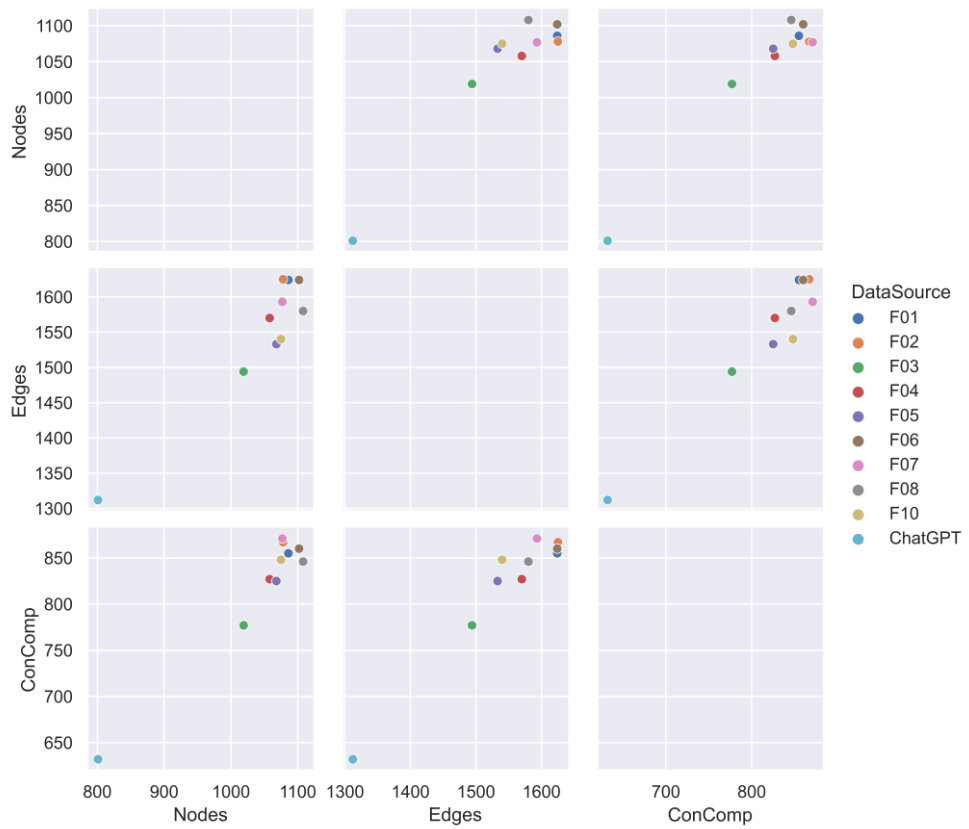


Figure 2. show the summary statistics generated from a pairplot where it shows the relationship among any pair in the plot. That includes Nodes, Edges, and Connected Components.

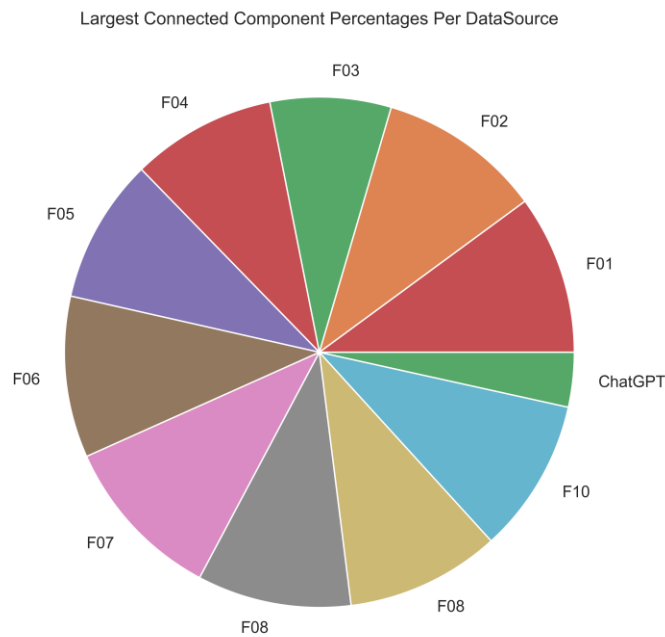


Figure 1 Figure 2 shows the size of the largest connected component in a network generated from the ChatGPT fake articles and compare it with the 10-fold. It is clear to visualize that the ChatGPT is the smallest among all the other slices.

3.3. The Classification of ChatGPT Fake Articles: Alzheimer's Disease Use Case

The previous analysis demonstrated how fake ChatGPT publications are fundamentally different in content and structure when compared with real publications. However, the above analysis has shown this using a 100 ChatGPT generated articles. Though, the analysis indeed succeeded to discriminate against those 100 generated publications, it does not solve the problem of detecting the case of one single fake article. In this section, we remedy this issue by introducing a modified version of the NeoNet algorithm¹⁴, which we now call xFakeBibs. The previous algorithm suffered from the lack calibration because of the unavailability decent fake publication dataset. Therefore, being able to identify lower and upper bounds acceptable for classification was not possible. We were able to get around this by introducing a support parameter which was tuned via experimentation according to source of datasets (social media, news, or scientific articles).

In this paper, we present the xFakeBibs algorithm that is specialized in classifying fake publications that are particularly generated using ChatGPT. We train the algorithm using a 100-publication partition selected from the Alzheimer's dataset from the TF-IDF bigrams generated earlier. The calibration a new process which makes this algorithm fundamentally different from NeoNet its predecessor. Algorithm 2 describes the steps that describe this process.

1. Using the training data, we constructed a network of bigrams to represent our training bigram network (same as before)
2. We extracted the largest connected components (LCC) and computed its size (same as before)
3. From each of the 10-fold (Fold1-Fold-10)
4. For each article, we extracted the TF-IDF bigrams
5. For each bigram that belongs to a given fold, we tested it the bigram contributed an edge to the LCC, and counted how many
6. After each fold being tested, we reset the training model to the original state
7. We computed the means of the number of edges that contributed to training for each fold
8. We designated the minimum value as the lower bound required number of bigrams that must be satisfied to label the article as real vs fake
9. We designated the maximum value as the upper bound required number of bigrams that must be satisfied to label the article as real vs fake
10. Using the lower-upper bound provided the necessary mechanism to
11. The final step was to analyze each ChatGPT fake article. If the number of bigrams contributed edges to the LCC training model, then it is

In Figure 4, we show a screenshot of the Python code that implemented the classification step for each of the ChatGPT individual article. By utilizing the lower and upper bounds, we measured the number of bigrams that contributed new edges to the LCC model without contributing new nodes. If it fell within the bounds, it is classified as POSITIVE, otherwise, it is classified as NEGATIVE.

```
## classify ChatGPT articles one at a time

lower_bound = 21.96
upper_bound = 24.93

counter = 0
positive = 0
negative = 0

for fake_article in stopped_chatGPT_test[0:100]:
    edges_added = measure_chat_GPT_article(fake_article, giant_cc)
    counter += 1
    if (len(edges_added) >= lower_bound) and (len(edges_added) <= upper_bound):
        positive += 1
        print('Article: ', counter, ', added edges: ', len(edges_added), 'POSITIVE')
    else:
        negative += 1
        print('Article: ', counter, ', added edges: ', len(edges_added), 'NEGATIVE')

print('\n\n-----\n')
print('NeoNet Classifier TRUE POSITIVE : ', format(positive/100), '.2f')
print('NeoNet Classifier FALSE NEGATIVE: ', format(negative/100), '.2f')
print('\n\n-----\n')
```

Figure 4. shows the classification step and how each article was detected as fake or real.

3.4. ChatGPT Article Classification Result

The average number of bigrams contributed between (21.96 – 25.12) number of edges to the LCC model. This set the lower bound of real articles to be 21.96 and the upper bound for 25.12. The classification results ended up with detecting 96 articles as NEGATIVE, while 4 articles have fallen into the acceptable lower and upper bounds. Table 3 shows the individual scores of each of those folds, which explains the lower/upper bounds of the classification step.

Table 3. shows the number of bigrams averages tht contributed edges to the LCC model, one per each fold.

FLD-1	FLD-2	FLD-3	FLD-4	FLD-5	FLD-6	FLD-7	FLD-8	FLD-9	FLD-10
24.93	25.12	21.57	23.80	22.49	22.91	23.52	22.65	23.73	21.96

To demonstrate how the averages have been indeed computed from analyzing the individual real publications, here we show Figure which displays the individual values of the bigrams generated from each article in all the 10 folds.

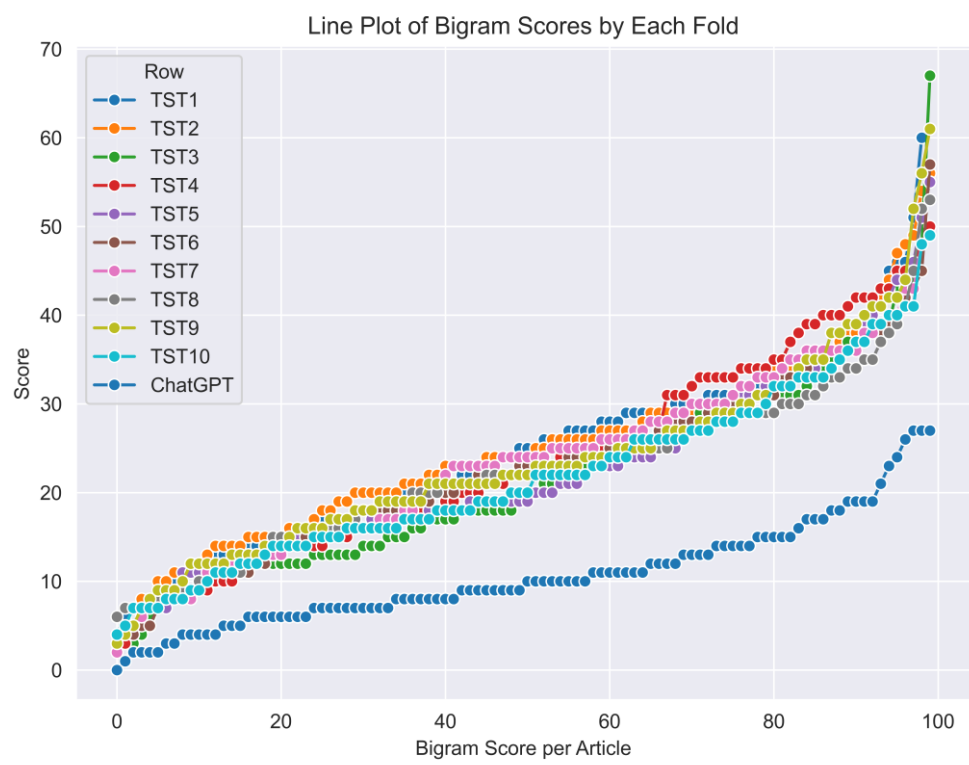


Figure 5. shows the bigram scoring behavior of the 10-folds of real publications being very similar, while the ChatGPT is exhibiting a completely different behavior than all other folds except for very few points that showed a similar behavior.

4. Discussion

In this work, we presented two different types of analysis which we used to evaluate fake publications that are generated by ChatGPT: (1) content analysis, and (2) a structural network analysis. Both analyses have shown that the publications generated by ChatGPT are significantly different from real publications produced by domain experts of the same subject. Though there was an overlap of some of the major bigram terms such as “Alzheimer’s Disease” or “AD disease”, the various aspects of the topic were explained in less clinical/scientific terms. As a result, this aspect made the contribution to the training model marginal when compared with terms that are derived from the real publications.

We also introduced a new version of the NeoNet algorithm, which we called xFakeBibs. The automatically generated content from ChatGPT provided a much-needed dataset which was missing previously. We empowered the xFakeBibs algorithm with a new step designed for calibration. Because of the trust we have in peer-reviewed publications, we used them to establish a calibration baseline. We used 10-fold chunks of publications to ensure no bias in the data. The 10 folds provided a lower/upper bound of what is expected of any publication to be classified as real. The algorithm detected 98 out of a 100 the publications generated by ChatGPT as NEGATIVE (not real publication). Although, this may seem to be a decent outcome of what an algorithm can do to detect fake publications. It remains concerning that the ChatGPT is intelligently capable of producing fake publications that would pass such tests. Fake publications pose great threat to our knowledge and safety.

Overall, we believe that the design of these types of algorithms is a step in the right direction. However, it is imperative to continue to advance such algorithms and methods to be able to safeguard our knowledge and fight Fake publications.

4.1. Principal Results

The most significant results of this research are as follows: (1) the medical publications generated by ChatGPT shows a significant behavior when analyzed for contents and data models (bigram networks). When working with a dataset of fake publications, such a behavior can be detected using proper computational methods and machine learning algorithms such those we present here, when trained and calibrated correctly, (2) despite the significant behavior exhibited by ChatGPT contents, it remains a challenging problem to detect a single publication as a ChatGPT-generated publication.

4.2. Limitations

We trained our algorithm using PubMed Abstracts which were extracted as a result of issuing the “Alzheimer’s Disease and Co-morbidities” query. Due to the lack of availability for Full Text papers, we have not trained the algorithm using Full-text.

4.3. Comparison with Prior Work

Various efforts have attempted the problem of detected fake news and publications. However, we believe this is the first effort to address assess this issue using the ChatGPT capability in generating contents such as publications. The assessment of the ChatGPT capabilities in generating publication is novel and we expect that this space will be rich in research and methods as time progresses

5. Conclusions

The ChatGPT is an artificial intelligence tool that has many impressive own capabilities. In fact, it inspired the creation of manuscript and provided a valuable dataset for experimentation, which was entirely missing before ChatGPT emerged. However, this is also a disturbing aspect that presents a threat to our knowledge and safety. Though, it is possible to detect fake science using machine learning algorithms and approaches. we have an ethical obligation to use such tools responsibly and regulate its uses. It is interesting to learn that some countries such as Italy have entirely banned ChatGPT. It is the opinion of the authors that such measures are too drastic, however, it is also not clear how such ethical issues are addressed. As ChatGPT stated: “It is up to individuals and organizations to use technology like mine in ways that promote positive outcomes and to minimize any potential negative impacts.”²⁹.

The future direction of this research is many: (1) extending the scope of PubMed Abstracts to Full-Text upon the availability, (2) Testing the algorithm with publications in multiple topics, (3) Computational fact -checking ChatGPT answers against well-known questions that require reasoning, (4) Training ChatGPT to answer domain specific questions (e.g., clinical, medical, chemical, and biological).

Acknowledgements: This publication is supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement Sano No 857533 and carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund.

Conflicts of Interest: None declared.

Abbreviations

JMIR: Journal of Medical Internet Research

RCT: randomized controlled trial

LCC: Largest Connected Component

TF-IDF: Term Frequency-Inverse Term Frequency

References

1. Synnestvedt MB, Chen C, Holmes JH. CiteSpace II: Visualization and Knowledge Discovery in Bibliographic Databases. *AMIA Annu Symp Proc.* 2005;2005:724-728.
2. Holzinger A, Ofner B, Stocker C, et al. On Graph Entropy Measures for Knowledge Discovery from Publication Network Data. In: Cuzzocrea A, Kittl C, Simos DE, Weippl E, Xu L, eds. *Availability, Reliability, and Security in Information Systems and HCI*. Lecture Notes in Computer Science. Springer; 2013:354-362. doi:10.1007/978-3-642-40511-2_25
3. Usai A, Pironti M, Mital M, Aouina Mejri C. Knowledge discovery out of text data: a systematic review via text mining. *J Knowl Manag.* 2018;22(7):1471-1488. doi:10.1108/JKM-11-2017-0517
4. Thaler AD, Shiffman D. Fish tales: Combating fake science in popular media. *Ocean Coast Manag.* 2015;115:88-91. doi:10.1016/j.ocecoaman.2015.04.005
5. Hopf H, Krief A, Mehta G, Matlin SA. Fake science and the knowledge crisis: ignorance can be fatal. *R Soc Open Sci.* 2019;6(5):190161. doi:10.1098/rsos.190161
6. Ho SS, Goh TJ, Leung YW. Let's nab fake science news: Predicting scientists' support for interventions using the influence of presumed media influence model. *Journalism.* 2022;23(4):910-928. doi:10.1177/1464884920937488
7. Frederickson RM, Herzog RW. Addressing the big business of fake science. *Mol Ther.* 2022;30(7):2390. doi:10.1016/j.ymthe.2022.06.001
8. Rocha YM, de Moura GA, Desidério GA, de Oliveira CH, Lourenço FD, de Figueiredo Nicolette LD. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *J Public Health.* Published online October 9, 2021. doi:10.1007/s10389-021-01658-z
9. Walter N, Brooks JJ, Saucier CJ, Suresh S. Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Commun.* 2021;36(13):1776-1784. doi:10.1080/10410236.2020.1794553
10. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum Behav.* 2021;5(3):337-348. doi:10.1038/s41562-021-01056-1
11. Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychol Sci Public Interest.* 2012;13(3):106-131. doi:10.1177/1529100612451018
12. Myers MG, Pineda D. Misinformation about Vaccines. In: *Vaccines for Biodefense and Emerging and Neglected Diseases*. Elsevier; 2009:255-270. doi:10.1016/B978-0-12-369408-9.00017-2
13. Matthews S. Government orders review into vitamin D role in Covid-19. Mail Online. Published June 17, 2020. Accessed April 13, 2023. <https://www.dailymail.co.uk/news/article-8432321/Government-orders-review-vitamin-D-role-Covid-19.html>
14. Abdeen MAR, Hamed AA, Wu X. Fighting the COVID-19 Infodemic in News Articles and False Publications: The NeoNet Text Classifier, a Supervised Machine Learning Algorithm. *Appl Sci.* 2021;11(16):7265. doi:10.3390/app11167265
15. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. *JMIR Med Educ.* 2023;9(1):e46885. doi:10.2196/46885
16. IEEE Special Issue on Education in the World of ChatGPT and other Generative AI | IEEE Education Society. Accessed April 13, 2023. <https://ieee-edusociety.org/ieee-special-issue-education-world-chatgpt-and-other-generative-ai>
17. Financial Innovation. SpringerOpen. Accessed April 13, 2023. <https://jfin-swufe.springeropen.com/special-issue---chatgpt-and-generative-ai-in-finance>

18. Languages. Accessed April 13, 2023.
https://www.mdpi.com/journal/languages/special_issues/K1Z08ODH6V
19. Do you allow the the use of ChatGPT or other generative language models and how should this be reported? JMIR Publications. Published March 16, 2023. Accessed April 13, 2023.
<https://support.jmir.org/hc/en-us/articles/13387268671771-Do-you-allow-the-the-use-of-ChatGPT-or-other-generative-language-models-and-how-should-this-be-reported->
20. Null N. The PNAS Journals Outline Their Policies for ChatGPT and Generative AI. Published online February 21, 2023. doi:10.1073/pnas-updates.2023-02-21
21. As scientists explore AI-written text, journals hammer out policies. Accessed April 13, 2023.
<https://www.science.org/content/article/scientists-explore-ai-written-text-journals-hammer-policies>
22. Fuster V, Bozkurt B, Chandrashekar Y, et al. JACC Journals' Pathway Forward With AI Tools. *J Am Coll Cardiol*. 2023;81(15):1543-1545. doi:10.1016/j.jacc.2023.02.030
23. Flanagan A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman "Authors" and Implications for the Integrity of Scientific Publication and Medical Knowledge. *JAMA*. 2023;329(8):637-639. doi:10.1001/jama.2023.1344
24. ChatGPT plugins. Accessed April 13, 2023. <https://openai.com/blog/chatgpt-plugins>
25. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*. 2023;9(1):e45312. doi:10.2196/45312
26. Tan CM, Wang YF, Lee CD. The use of bigrams to enhance text categorization. *Inf Process Manag*. 2002;38(4):529-546. doi:10.1016/S0306-4573(01)00045-0
27. Hirst G, Feiguina O. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Lit Linguist Comput*. 2007;22(4):405-417. doi:10.1093/lit/fqm023
28. Hamed AA, Ayer AA, Clark EM, Irons EA, Taylor GT, Zia A. Measuring climate change on Twitter using Google's algorithm: perception and events. *Int J Web Inf Syst*. 2015;11(4):527-544. doi:10.1108/IJWIS-08-2015-0025
29. ChatGPT. Accessed April 12, 2023. <https://chat.openai.com>