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:orresponding author: 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 new algorithm, called xFakeBibs, we show the significant difference between ChatGPT-generated fake publications and real publications. Specifically, we triggered ChatGPT to generate 100 publications that were related to Alzheimer's disease and comorbidity. Using TF-IDF, using the real publications, we constructed a training network of bigrams comprised of 100 publications. Using 10-folds of 100 publications each, we also 10 calibrating networks to derive lower/upper bounds for classifying articles as real or fake. The final step was to test xFakeBibs against each of the ChatGPT-generated articles and predict its class. The algorithm successfully assigned the POSITIVE label for real ones and NEGATIVE for fake ones. **Results:** When comparing the bigrams of the training set against all the other 10 calibrating folds, we found that the similarities fluctuated between (19%-21%). On the other hand, the mere bigram similarity from the ChatGPT was only (8%). Additionally, when testing how the various bigrams generated from the calibrating 10-folds against ChatGPT we found that all 10 calibrating folds contributed (51%-70%) of new bigrams, while ChatGPT contributed only 23%, which is less than 50% of any of the other 10 calibrating folds. The final classification results using the xFakeBibs set a lower/upper bound of (21.96-24.93) number of new edges to the training mode without contributing new nodes. Using this calibration range, the algorithm predicted 98 of the 100 publications as fake, while 2 articles failed the test and were classified as real publications. **Conclusions:**This work provided clear evidence of how to distinguish, in bulk ChatGPT-generated fake publications from real publications. Also, we also introduced an algorithmic approach that detected fake articles with a high degree of accuracy. However, it remains challenging to detect all fake records. ChatGPT may seem to be a useful tool, but it certainly presents a threat to our authentic knowledge 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 a camp of optimism that sees potential and seeks to utilize it. On the other hand, there are doubters who remain skeptical and search for validation and further assessments to decide how this tool affects our lives. This split provided a strong motivation for this work and triggered the effort of providing an assessment tool of fake publications generated by ChatGPT. Without any

 doi:10.20944/preprints202304.0350.v2

doubt, real science (and publications) is one of the most sacred sources of knowledge. That is because it is an invaluable source of any future discovery [1–3]. The spread of various predatory journals contributed to fake science and caused to be on the rise[4–7]. 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 putting others at harm[10–12]. Another disturbing example that occurred also during the pandemic was the propagation of a article that reported fake results about how the deficiency of vitamin D led to the death of 99 per of the population studied. Though the article was eventually withdrawn, the damage of 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 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 Slack[24], (4) the development of educational tools (e.g., Wolfram); and potential of developing learning and educational support tools (e.g., Medical Licensing Examination[25].)

2. Methods

2.1. Data Collection

We used two different datasets: (1) for training and calibrating the algorithm, we queried PubMed for "Alzheimer's Disease and comorbidities" and collected 1000 abstracts of publications that are human produced, (2) for detecting ChatGPT articles, we asked ChatGPT to generate 100 abstracts that are also related to "Alzheimer's disease and comorbidities". Both datasets were preprocessed using the same mechanism to address noise and stop words in the data.

Statistical Analysis

We performed two different types of analysis to discriminate contents of real publications against contents of ChatGPT: (1) using an equal number of records, we compared the Term Frequency-Inverse Document Frequency (TFIDF)[26–31] of bigrams generated from the two sources (ChatGPT records, PubMed records), (2) The statistical structural analysis of networks from both sources.

2.1.1 ChatGPT Bigram-Similarity Analysis

Bigrams are any two consecutive words that may prove significant in any given text-based dataset [32,33]. Using the Term Frequency-Inverse Term Frequency (TF-IDF) we generated bigrams from three different types of resources: (1) a training dataset, which is a slide of a 100 real publications, (2) a calibration using 10-folds each of which is also 100 publications, and (3) the classification of 100 publications generated using the ChatGPT. We compared the TF-IDF scores of the training dataset with all the calibrating 10-fold to establish lower/upper bounds test against the ChatGPT fake articles. We computed the bigram similarities (ones that overlapping in both the training and against the calibration). The process was repeated once more to test bigram similarities generated from ChatGPT content against the 10-folds calibrations. The similarity ratios from the two comparisons offered significant difference between bigrams of real publications and bigrams of ChatGPT-generated publications. Figure shows a side-by-side WordCloud weights TF-IDF

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

bigrams, the one to the left is from real publications, while the one to the right was generated from ChatGPT fake records. While "Alzheimer's disease" and "cognitive impairment" are common in both real publications and ChatGPT records, they are weighted and ranked differently. Table 1 shows the scoring of top-4 with similarity of color-coded bigrams:

Table 1. shows the tops-5 ranked bigrams in ChatGPT vs Real Publcations, with similar color-coded when overlapping.

	Bigram1	Bigram2	Bigram3	Bigram4	Bigram5
ChatGPT	Ad patients	Cognitive impairment	older adults	Alzheimer's disease	Risk factors
Real Bibs	Alzheimers disease	Disease ad	Ad patients	Cognitive impairment	Increased risk

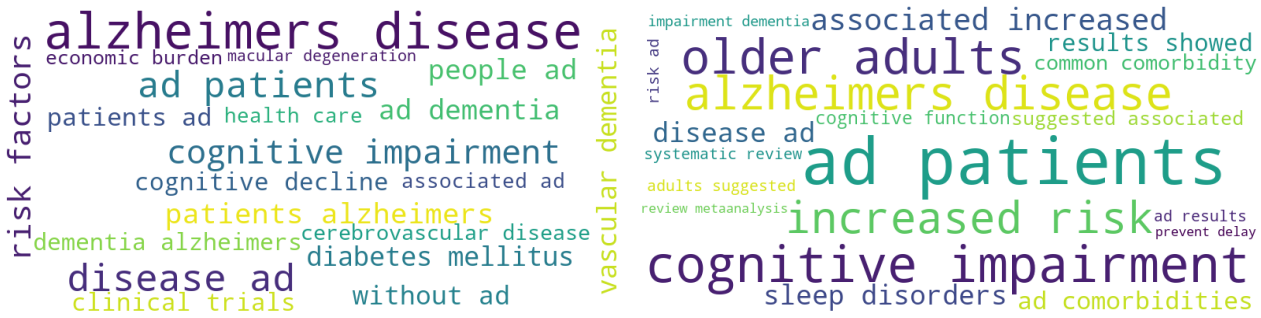

Figure 1. shows side-by-side WordCloud Compairson of the top-20 bigrams of real publications vs ChatGPT.

2.1.2. ChatGPT-Network Structural Analysis

In addition to the bigram similarity analysis above, we also utilized the natural mechanism for constructing networks of two connected word [14,34]. When compiled together, each word can be connected to multiple other words, which form a network of words where the semantic of each edge is a bigram predicate.

The purpose of this type of 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 two networks, namely, the largest connected components (i.e., the largest number of words that made up the bigrams). Specifically, we computed the largest connected component of the training dataset to establish a baseline of comparison. We then proceeded by computing the largest components of all the 10-folds (one-fold at a time) to establish calibration. The final step was to compute the largest connected component from the ChatGPT network of bigrams to test against the calibration. Algorithm 1 describes the calibration step-by-step.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

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:

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

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

Algorithm 1 describes the calibrations steps computed using 10 individual folds to be compared agasint the training baseline

3. Results

3.1. Outcome of ChatGPT Content Analysis: The Statistics of Bigrams Comparison

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

Table 2. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

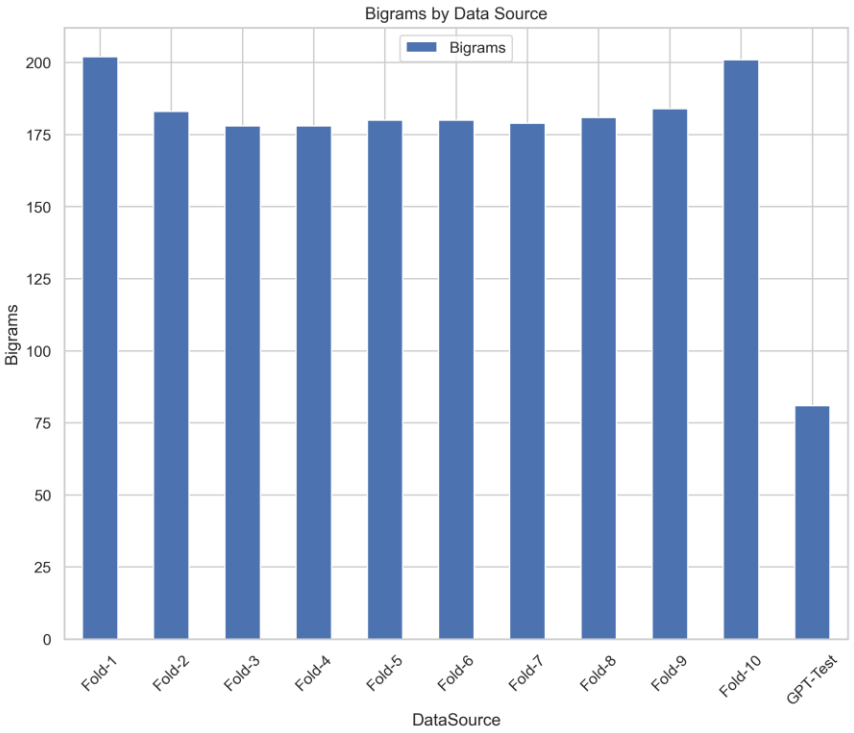

Figure 2. shows the number of overlapping bigrams of 10-folds of publcations vs ChatGPT generated publications, when compared with a training dataset from real publications. .

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%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

Table 3. 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%).

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

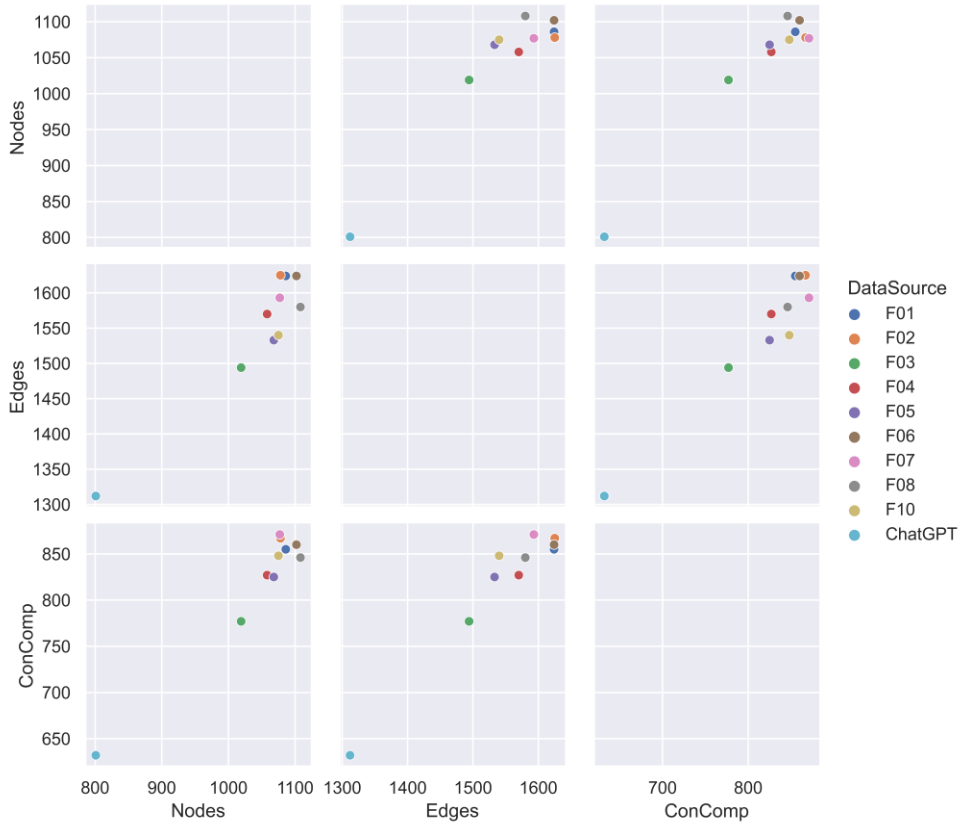

Figure 3. show the summary statistics generated from a pairplot where it shows the relationship among any pair in the plot. That includes Nodes, Edges, and Connected Components.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

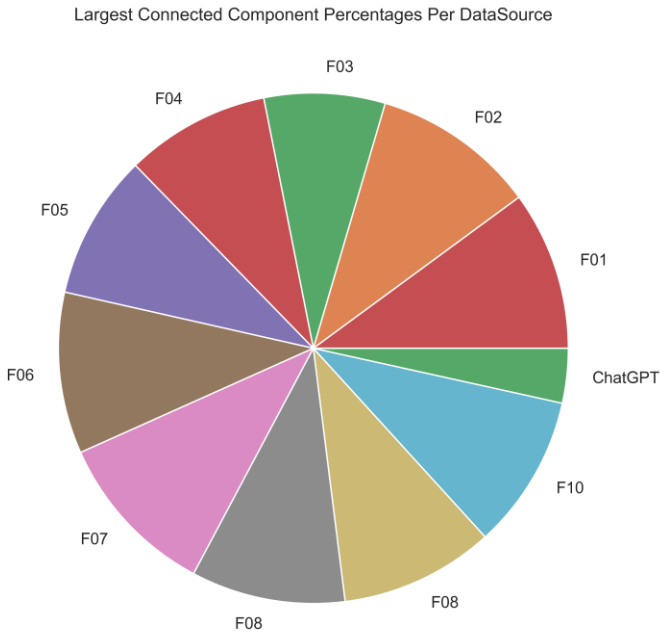

Figure 4. Figure 2 shows the size of the largest connected component in a network generated from the ChatGPT fake articles and compare it with the 10-fold. It is clear to visualize that the ChatGPT is the smallest among all the other slices.

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 algorithm[14], 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

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 classified as POSITIVE (a real publication), otherwise, it is classified as NEGATIVE (fake publication)

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

```
## 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 5. shows the classification step and how each article was detected as fake or real.

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 4. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 April 2023 doi:10.20944/preprints202304.0350.v2

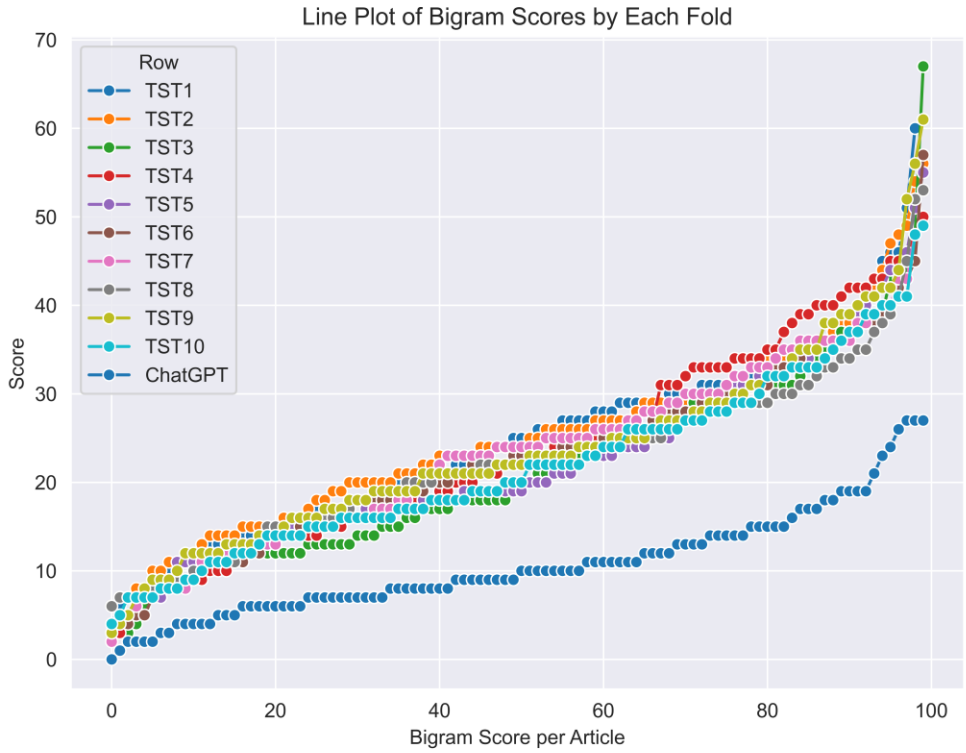

Figure 6. shows the bigram scoring behavior of the 10-folds of real publications being very similar, while the ChatGPT is exhibiting a completely different behavior than all other folds except for very few points that showed a similar behavior.

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

 doi:10.20944/preprints202304.0350.v2

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 and calibrated our algorithm using PubMed Abstracts, which were extracted as a result of issuing the "Alzheimer's Disease and Co-morbidities" query. This is because ChatGPT failed to generate a large dataset without halting. The simplification of using abstracts was a work around.

4.3. Comparison with Prior Work

Various efforts have attempted the problem of detecting fake news and publications. However, we believe this is the first effort to address this issue using 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

When I asked a highschooler what she knows about ChatGPT the answer was "*Do you mean that thing that does your homework for you?*". Indeed, ChatGPT is an a highly intelligent tool that has many impressive own capabilities. In fact, it provided a valuable dataset for experimentation, which was entirely missing before ChatGPT emerged. However, it is also a disturbing aspect that threatens the future of our science if the younger generations, the pioneers of the future, use it to plagiarize. Though, it is possible to detect fake science using machine learning algorithms, we have an ethical obligation to use such a tool 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."[35].

The future directions of this research are many: (1) Using the ChatGPT APIs to generate full publications; and compare with full-text archived articles, (2) Testing the algorithm with publications in multiple topics, (3) Fact-checking ChatGPT answers for 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. The authors would like to thank Dr. Marian Bubak for inspiring and supporting the assessment of ChatGPT and its impact on our knowledge. The authors also acknowledge Laila Hamed for her valuable perspective on ChatGPT.

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 Nicolete 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, Chandrashekhar 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. Flanagin 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. Aizawa A. An information-theoretic perspective of tf–idf measures. *Inf Process Manag*. 2003;39(1):45-65. doi:10.1016/S0306-4573(02)00021-3
27. Qaiser S, Ali R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *Int J Comput Appl*. 2018;181. doi:10.5120/ijca2018917395
28. Ramos J. Using TF-IDF to Determine Word Relevance in Document Queries.
29. Trstenjak B, Mikac S, Donko D. KNN with TF-IDF based Framework for Text Categorization. *Procedia Eng*. 2014;69:1356-1364. doi:10.1016/j.proeng.2014.03.129
30. Wu HC, Luk RWP, Wong KF, Kwok KL. Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans Inf Syst*. 2008;26(3):13:1-13:37. doi:10.1145/1361684.1361686

 doi:10.20944/preprints202304.0350.v2

31. Zhang W, Yoshida T, Tang X. A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Syst Appl*. 2011;38(3):2758-2765. doi:10.1016/j.eswa.2010.08.066
32. 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
33. 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/llc/fqm023
34. 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
35. ChatGPT. Accessed April 12, 2023. https://chat.openai.com