ReproHum #0866-04: Another Evaluation of Readers' Reactions to News Headlines

Zola Mahlaza[1], **Toky Raboanary**[1], **Kyle Seakgwa**[1,2], **C. Maria Keet**[1]
[1]University of Cape Town, Cape Town, South Africa
[2]University of the Western Cape, Cape Town, South Africa
{zmahlaza,traboanary,mkeet}@cs.uct.ac.za, SKGKYL001@myuct.ac.za

Abstract

The reproduction of Natural Language Processing (NLP) studies is important in establishing their reliability. Nonetheless, many papers in NLP have never been reproduced. This paper presents a reproduction of Gabriel et al. (2022)'s work to establish the extent to which their findings, pertaining to the utility of large language models (T5 and GPT2) to automatically generate writer's intents when given headlines to curb misinformation, can be confirmed. Our results show no evidence to support two of their four findings and they partially support the rest of the original findings. Specifically, while we confirmed that all the models are judged to be capable of influencing readers' trust or distrust, there was a difference in T5's capability to reduce trust. Our results show that its generations are more likely to have greater influence in reducing trust while Gabriel et al. (2022) found more cases where they had no impact at all. In addition, most of the model generations are considered socially acceptable only if we relax the criteria for determining a majority to mean more than chance rather than the apparent $> 70\%$ of the original study. Overall, while they found that "machine-generated MRF implications alongside news headlines to readers can increase their trust in real news while decreasing their trust in misinformation", we found that they are more likely to decrease trust in both cases vs. having no impact at all.

Keywords: text generation, reproduction, misinformation

1. Introduction

The reproduction of Natural Language Processing (NLP) studies is critical in establishing the reliability of published findings. This is especially timely since there is evidence that a number of NLP studies are not repeatable (Belz et al., 2023). The low levels of replicability observed in these investigations warrants significant attention given that in many other fields, such as the social and medical sciences, low levels of replicability have also been observed in large scale replication efforts (OpenScienceCollaboration, 2015). The results in such studies triggered a decade-long reckoning with this "reproducibilty crisis" (Baker, 2016). It eventually led to more stringent standards being adopted for reporting results, a push toward preregistered research designs, and the adoption of more open science methods, like the sharing of datasets (Vazire, 2018).

While the low levels of replicability initially called into question the reliability of results in the social and medical sciences, efforts to address these shortcomings triggered what has been called a "credibility revolution" due to widespread adoption of the aforementioned improvements (Vazire, 2018). For the NLP community to undergo a similar "credibility revolution", more research like (Belz et al., 2023) needs to be undertaken to ascertain the extent of its reproducibility problem. As part of an effort to ascertain the extent to which existing work is reproducible (Belz and Thomson, 2024), this paper reports on the reproducibility of the human evaluation study conducted by Gabriel et al. (2022).

The work by Gabriel et al. (2022) focuses on investigating the utility of text generation models for automatically generating a writer's intent when given a news headline, as a means of combating misinformation. While the original work focuses on numerous tasks (e.g., it described the creation of a misinformation news headline corpus with human annotations of the writer's intent, readers' perception, possible actions that could be taken by the reader, and the likelihood of spread of the associated article), our sole focus in on the reproducibility of its human evaluations.

We investigate the reproducibility of the original study via a survey with 42 crowd-workers[1] who are based in the United States and judge the headline and intent pairs from the original study. The nature of the study is kept the same, where possible, and we compare the resulting findings to establish whether there is any difference with the original work. We have found that the results obtained with our survey contradict 2/4 of the findings from the original study and we can partially support two of the original study's findings. Specifically, with respect to the partially supported findings, most of the models' generations are considered socially acceptable if the criteria for determining a majority means more than chance[2] instead of ≥ 70", a value

[1]One was excluded in the final analysis as they submitted incomplete survey responses

[2]We assume that "chance" means 50%

Proceedings of the HumEval @LREC-COLING 2024, pages 274–280
Turin, Italy. May 21st, 2024.

that can be inferred from the results. In addition, while all models were rated as being capable of influencing readers to trust or distrust, T5's generations are more likely to have greater influence in reducing trust while Gabriel et al. (2022) found more cases where they had no impact at all.

The rest of the paper is structured as follows. Section 2 summarises how the original study was conducted and lists the findings that emanated from it. Section 3 describes how the reproduction survey was set up, methods used to compare Gabriel et al. (2022)'s work with the current study, and the results we obtained in our survey. The differences and similarities with respect to findings between the two studies are discussed in Section 4, and Section 5 concludes.

2. Original study

While most work on combating misinformation pursues the creation of models to classify headlines, or articles, as being real or misinformation, Gabriel et al. (2022) takes a different approach towards building AI models. They investigate the extent to which machine-inferred writer's intents can improve reader's ability to identify misinformation. They do so by creating a human-annotated news corpus of headlines and intents with which they fine-tune pre-trained language models. The utility of the generated intents is evaluated by humans.

2.1. Dataset and models

The headlines were sourced from published misinformation datasets about Covid-19 (Cui and Lee, 2020; Gruppi et al., 2021; Network, 2024; Shapiro et al., 2020), climate change (Gruppi et al., 2021; Nørregaard et al., 2019), and cancer (Cui et al., 2020). The authors use the dataset to train models to automatically generate the writer's intent when given a headline and associated information (e.g., domain of the associated article/headline — either Covid-19, climate change, or cancer). The writers' intents are generated using two pre-trained language models, namely T5 (Raffel et al., 2020) and GPT2 (Radford et al., 2019).

2.2. Evaluation

The corpus and models are used to (1) investigate whether the headlines are trustworthy, be this with or without the writers' intent annotations that are automatically generated by the models, (2) determine whether the generated writers' intents are coherent and relevant, (3) establish whether the writers' intents are socially acceptable, and (4) ascertain whether the headlines and/or writers' intents perpetuate negative social biases or stereotypes.

2.3. Findings

Mechanical Turk (MTurk) workers' judgements of the trustworthiness of each news headline (with and without the intent) shows that while there were changes after seeing the intent, in the best case (i.e., intents generated by T5) there was only a weak positive correlation with the true class label (i.e., real/misinformation). Workers were also asked to judge the overall quality of the machine-generated intent in terms of coherence and its relevance to the headline on a 5-point Likert scale. The judgements show that the intents generated by T5 were perceived to have better quality, with an average of 3.74. Workers were also asked to judge whether the writer's intent conveys feelings or thoughts that are socially acceptable on a binary scale. In the case of one of the T5 variants, a model with the highest socially acceptable intents, we see a percentage of 75.30%. While worker's judgements of the capability of the beliefs and/or news events to perpetuate negative social biases or stereotypes were solicited, the results are not reported.

3. Reproducibility Study Design

The goal of our study was to reproduce the human evaluations using the same resources and methods as the original study, where possible. We did not aim to recreate their text generation models from scratch, but only reproduce the human evaluation thereof. We conducted the survey via Prolific[3] and an institutionally hosted version of LimeSurvey[4]. The human evaluation datasheet (Shimorina and Belz, 2022) for the study is shared via Github[5].

3.1. Survey

We created a survey using a dataset of 600 tuples of human authored headlines and automatically generated writers' intents. The dataset was sourced from Gabriel et al. (2022) via the organisers of the ReproNLP (Belz and Thomson, 2024) task. Each writer's intent is either 'real' or 'misinformation' and it is generated by one of the two types of models described in Section 2. Since the original study does not specify the number of texts evaluated by each participant, we split the dataset into 13 batches of 45 headline and intent pairs and one batch with 15 pairs. This was done to prevent collecting low quality judgements due to participant fatigue. Each batch was packaged into a survey where the participant is first given instructions, verbatim from the

[3] https://www.prolific.com/

[4] https://survey.cs.uct.ac.za/limesurvey/

[5] https://github.com/nlp-heds/repronlp2024

original study, describing what to expect as part of the survey ("You will read a sentence fragment describing a belief someone reading a news headline would have...") and what questions will be posed (e.g., "Please rate the quality of the belief description based on the following questions..."). They are then asked to judge quality of headline and intents, as shown in Figure 1.

Since the original study elicited 3 unique judgements per headline, we attempted to abide by the criteria as much as possible. We created a Python application (a web application created using the Django framework) to randomly assign a Prolific worker to one of the 14 surveys, provided it has less than three responses at the time of initiating the task. The survey was distributed to 42 Prolific participants that are based in the US, have 99% task approval, and have at least 200 tasks that have been approved.

Evaluation strategy There are two components to the evaluation. First, the calculations as by Gabriel et al. (2022). Overall Quality (coherence and relevance) is recorded on a 1-5 Likert scale. Influence on Trust is measured as more (+) or less (-) trustworthy, calculated as percentages, based on a 5-point scale that asked for the readers' perception. Third, for the sociopolitical acceptability, participants rate "their perception of the beliefs invoked by an implication in terms of whether they represent a majority (mainstream) or minority (fringe) viewpoint", where Gabriel et al "refer to "minority" viewpoint broadly in terms of less frequently adopted or extreme social beliefs, rather than in terms of viewpoints held by historically marginalized groups". This is reported as a percentage. We also recorded, on a nominal scale, the capacity of the headline and/or intent to perpetuate negative social biases or stereotypes. We report this as a percentage, even though it is not reported in the original study.

While the methods in Gabriel et al. (2022) do not describe further details, the results table indicates also "Corr w/ Label (all gens)", "Corr w/ Label (quality ≥ 3)", and statistical significance. We take these to be correlations and a student-t test (with "$p < .05$").

Second, the comparisons of the results obtained in this reproduction are to be compared to the original results as reported in Gabriel et al. (2022). This involves both a numerical comparison and whether the same conclusions can be drawn from the results obtained. We first established where there is a difference in the computed percentages via a two-sample proportion hypothesis test (i.e., Z-test). We do not compare whether there is a significant difference between the means since they are only computed for Likert scales and are likely to lead to misinterpretations, especially since there is a potential difference in how the evaluated data was batched.

Following that, we guided our comparisons using the findings (abbreviated **F** in the list) of the original study:

F1: "The T5-large model was rated as having slightly higher quality generations than the other model variants": We compared whether T5's average score was higher than the alternative model.

F2: "Most model generations were rated as being "socially acceptable"": We calculated whether most generations were judged as being socially acceptable. The original study does not specify what they deem a 'reasonable' majority is, so the cutoff point could belong to any value in the range (51-100), but we decided to use $\geq$70.

F3: "All models were rated capable of influencing readers to trust or distrust": We counted the number of models for which the change in trust of distrust is non-zero.

F4: "[For] T5-base [, there is a] consistent correlation between the actual label and shifts in trustworthiness scores before and after annotators see the generated writer's intent": Working under the assumption that the original study established the consistency of a positive and significant correlation for T5-base between two binary attributes (shift in trust & label), we calculated the tetrachoric correlations[6] for the models and tested whether they are positive and significant[7] (two-tailed t-test, alpha=0.05) for T5-base but not the other models.

3.2. Results

The fourteen batches received a variable number of responses, as shown in Table 1. The results presented henceforth are not relevant to batch 10 since it did not get any responses. We also excluded a single response in batch 7 since one participant submitted incomplete responses.

Results of the participants' judgements, from the original and current study, are presented in Table 2. The results pertaining to the capacity of the intent and/or headline to perpetuate negative social biases or stereotypes are given in Table 3.

A comparison of the percentage of headlines that are considered socially acceptable, via the Z-test, shows that there is a significant difference (z = 2.51 for T5-base, z = 2.39 for T5-large, and z = 3.48

[6]Calculated using https://py-pair.readthedocs.io/

[7]H0: there is no linear relationship between the shift in trust and headline label.

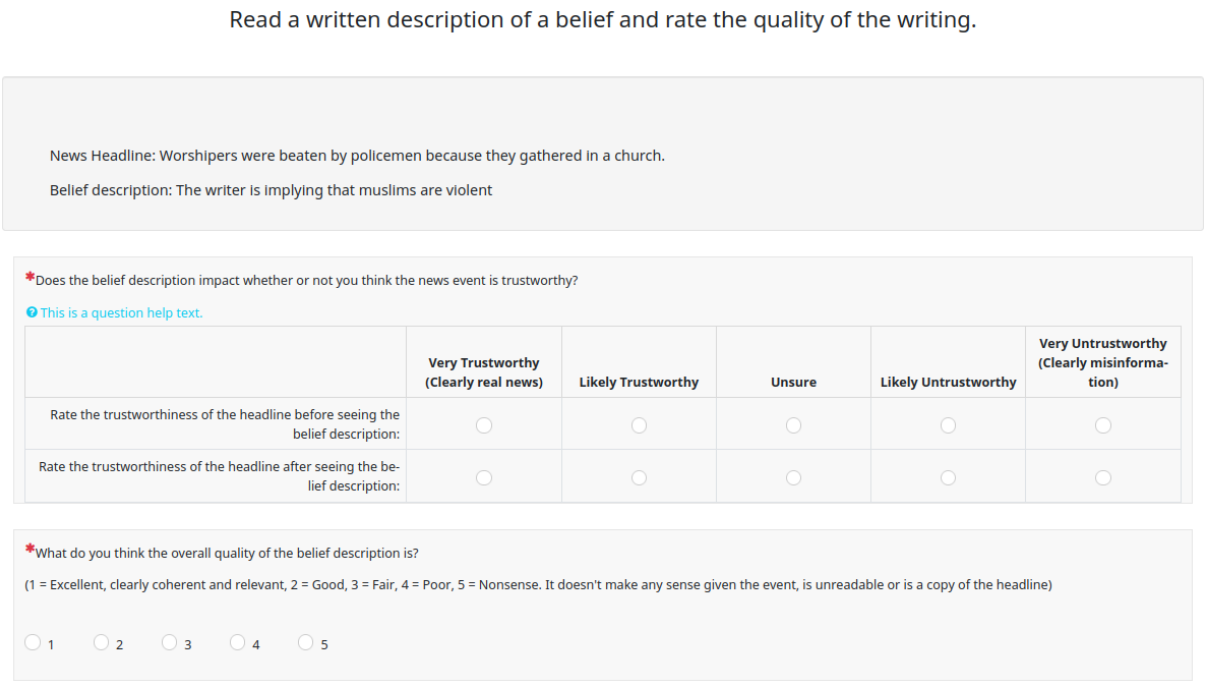

Figure 1: Screenshot of a section of the survey, with one of the news headlines, its belief description, and two of the questions for that headline and belief description.

Table 1: Number of responses for each batch. One of the responses from batch 7 was excluded because the crowdworker had not answered all the questions.

Batch	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Responses (#)	8	2	2	2	3	2	5*	2	2	0	3	2	3	5

for GPT2-large vs. critical value of 1.96) between the original study and our work for all three models. Specifically, the original study reported higher values of socially acceptable headlines. The extent to which the measured scores of social acceptability differ is not the same for all the models as Belz (2022)'s coefficient of correlation, given in Table 4, shows that GPT2-large exhibits the worst reproducibility while T5-large is better than the two alternative models.

Using the same test, we established that there is also a significant difference (z = -4.01 for T5-base, z = -3.80 for T5-large, and z = -1.09 for GPT2-large vs. critical value of 1.96) in the percentage of texts where there is an increase in trust after seeing the intent in the case of T5. However, we found no evidence that there is a significant difference in the case of GPT2. Noteworthy is that the original study recorded GPT2 as the model for whose intents have the greatest capacity to increase trust in the headline while the opposite was true in the current study (even if not statistically significant). The Z-test also showed that there is a significant difference (z = -40.09 for T5-base, z = -39.33 for T5-large, and z = -40.11 for GPT2-large vs. critical value of 1.96) in the percentage of texts for which there was decrease in trust after seeing the intents.

The differences, with respect to a shift in trust, can be attributed to the high number of intents/headlines for which there was no change in participants' trust in the original study whereas no participant's trust was unaffected in our study.

We have found statistically significant evidence that there is no correlation between shift in trust and the class label in the case of T5-base, unless we exclude low quality (i.e., quality < 3) generations.

4. Discussion

We first compare our results to those reported in Gabriel et al.'s paper and then reflect on the reproducibility process.

4.1. Comparison of results with the original study

We now turn to confirm whether our study was able to confirm Gabriel et al. (2022)'s original four findings, as described in Section 3.1:

F1: Our results contradict this finding. Specifically, we found that GPT2, the alternative model, had higher quality generations than T5 (both

Table 2: Human evaluations from the original and reproduced study. Cells from original study marked with * indicate the statically significant existence of a correlation for $\alpha = 0.05$. Cell marked with ‡ indicate a statically significant the lack of a correlation for the same α value. Abbreviations: Orig = Original (i.e., Gabriel et al. (2022)), Corr = Correlation

			Influence				
	Model	**Quality** (1-5)	**+Trust** (%)	**-Trust** (%)	**Corr. (all gens)**	**Corr. (quality ≥ 3)**	**Socially accept.** (%)
Orig.	T5-base	3.61	8.33	7.82	**0.24***	**0.30***	**75.30**
	T5-large	**3.74**	7.73	9.76	-0.03	0.09	74.66
	GPT2-large	3.46	**9.70**	**13.10**	-0.04	0.10	74.66
Ours	T5-base	2.61	**16.03**	83.97	0.07‡	**0.99**	**68.67**
	T5-large	2.56	14.77	85.43	**0.99**	**0.99**	68.31
	GPT2-large	**2.77**	11.68	**89.38**	**0.99**	**0.99**	65.30

Table 3: Percentage of headlines and intents that perpetuate negative stereotypes. Abbreviations: Sent = Sentence

Model	**Both do**	**Neither do**	**Sent.**	**News event**
T5-base	13.27	74.16	5.84	6.73
T5-large	18.4	73.09	7.96	5.49
GPT2-large	14.51	68.85	11.5	5.13

Table 4: Precision results for the socially acceptable attribute between the original and current study. Abbreviations: Unb. stdev = unbiased standard deviation, CV* = Belz (2022)'s coefficient of variation

Model	**Mean**	**Unb. stdev**	**CV***
T5-base	71.985	5.876	9.1827
T5-large	71.485	5.628	8.8564
GPT2-large	69.980	8.295	13.3352

T5-base and T5-large). GPT2 also had the highest CV*, denoting the poor reproducibility when compared to the alternative models. In addition, the quality judgements of all the models were lower by about 1 point on a 5-point Likert scale.

F2: It is not clear what value is used by Gabriel et al. (2022) to determine a majority and our interpretation of their results suggests that they used ≥ 70. Based on that interpretation, our results do not support this finding as we had fewer generations that are socially acceptable. There is a difference of 6-9 percentage points between our results and theirs and the difference is statistically significant. This finding can only be supported to relax the cut-off point from ≥ 70 to ≥ 51.

F3: Our results confirm that all the models are judged to be capable of influencing readers' trust or distrust. However, there was a significant difference in T5's capability to reduce trust. Specifically, our results show that its generations are more likely to have greater influence in reducing trust while Gabriel et al. (2022) found more cases where they had no impact at all. There were also more cases where T5-base's generations positively shifted trust vs. GPT2.

F4: Our results contradict this finding. In fact, they show that T5-base is the only model for which there is no consistent correlation between the actual label and shifts in trustworthiness scores. When low quality (i.e., quality < 3) generations are included, our evidence demonstrates that there is no linear relation between the attributes. T5-large and GPT2 are the only models for which there is a consistent and strong correlation.

4.2. Challenges reproducing the study

The experiment's methods could not be reproduced exactly, due to several reasons. First, the ReproNLP project (Belz and Thomson, 2024) moved to Prolific (cf. MTurk in the original study). This had consequences for technically setting up the task. We cannot directly use the form created and used by Gabriel et al. (2022) for MTurk in Prolific. We had to use experimental software compatible with Prolific, such as LimeSurvey and Gorilla, or develop a new form hosted on another server and link it to Prolific to create a similar survey. We decided to use LimeSurvey to facilitate the experiment while maintaining the same objective. This allowed us to rely on its existing features such as recording

responses and collection of metadata pertaining to submission times, time taken to complete the survey, etc. However, the layout differs from Gabriel et al. (2022)'s survey.

Second, it was our decision to break up the task into batches, which may, or may not, have been done in the original study, as described in Section 3. With an estimated task completion time of 45 minutes for a batch of 45 headlines, it was deemed unreasonable to make a participant assess a batch of 600 headlines since that would have taken approximately 10 hours.

The change in evaluation platform used also resulted in a difference in the number of participants who evaluated each headline/intent pair per survey between the two studies. While our evaluation instrument was set up to abide by the upper limit of 3 responses per survey as much as possible, as mentioned in Gabriel et al. (2022), via assigning each worker to a batch that did not have $\geq$3 responses already. We still obtained more than three responses for some batches, as included in Table 1, since there were cases where some workers were assigned to batches for which there were other users who were already evaluating but had not submitted their responses.

It is also possible that the way the headline and intent are shown to each participant may differ from the original study. We had access to a screenshot of the instrument used by Gabriel et al. (2022) and we determined that it uses a template (i.e., "News Headline: $[sentence]") at the top of the survey to display the information to be evaluated. However, it was unclear how and when the intent was presented to each participant. As such, we took the decision to include the intent alongside the news headline, as shown in Figure 1.

The instrument used by Gabriel et al. (2022) solicited judgements on a nominal scale, to determine whether the writer's intent and/or headline perpetuates negative social biases or stereotypes. However, the solicitation of those judgements was not described in the publication nor was there a presentation of the associated results. Nonetheless, we decided to collect and report such judgements for completeness.

An average cost of 12.60GBP (max of 12.62GBP) was spent on renumerating each participant and the figure includes the Prolific service fee of 3GBP and a value added tax between 0.59-0.61GBP. This was in line with the ReproNLP task's (Belz and Thomson, 2024) instructions which mandated a value of 12GBP per hour. The extent to which this figure differs from the original study is unclear since the original study does not specify how much evaluators were compensated. It only mentions the workers were paid $.6 per human intelligence task. However, it is unclear how many items were evaluated by each participant. In addition, of the 600 items to be evaluated, the authors excluded 12 since they were deemed as malformed or unsuitable, but the exclusion criteria were not reported/communicated.

Gabriel et al. (2022) note that they "obtained an Institutional Review Board (IRB) exemption for annotation work, and ensured annotators were fairly paid given time estimations.", however it is unclear if ethics approval was obtained for the evaluation task. It is possible that the authors either use the same term "annotator" to refer to both annotators and evaluators, when discussing ethics approval, since the latter are a subset of the crowd-workers that were initially recruited, or they may have deemed it unnecessary to seek ethics approval for the evaluation. The University of Cape Town does require ethics approval for experiments involving humans or large datasets, however, and thus needed to be obtained from the Science Faculty Ethics Committee. Besides filling in the form, this involved writing from scratch a research proposal, a data management plan, and a task-adapted consent form. It was approved with reference number SCI/00635/2024.

5. Conclusion

Our reproduction of the human evaluation component of the research reported in Gabriel et al. (2022) contradicts two out of the four findings reported in Gabriel et al. (2022), being **F1** and **F4**. Specifically, GPT2's generations had better quality even though Gabriel et al. (2022) found that T5's generations are better (**F1**) and T5 did *not* exhibit consistent correlations between the actual label and shifts in trustworthiness scores due to the inclusion of low quality (i.e., quality < 3) intents (**F4**). Most model generations were not rated as being "socially acceptable" (**F2**), unless one were to lower the cut-off point for determining a majority from $\geq 70\%$ to $\geq 51\%$. Lastly, our results confirm that all the models were capable of influencing readers' trust or distrust (**F3**). However, unlike the original study, T5's generations are more likely to have greater influence in reducing trust even though Gabriel et al. (2022) found more cases where they had no impact at all and there were more cases where T5-base's generations positively shifted trust vs. GPT2.

6. Acknowledgements

The work was in part funded by EPSRC Grant No. EP/V05645X/1 for the ReproHum project, and the payment of the crowdworkers specifically.

7. Bibliographical References

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:7604:452—4:454.

Anya Belz. 2022. A Metrological Perspective on Reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz and Craig Thomson. 2024. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3676–3687. Association for Computational Linguistics.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 492–502, New York, NY, USA. Association for Computing Machinery.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3108–3127. Association for Computational Linguistics.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2021. Nela-gt-2020: A large multi-labelled news dataset for the study of misinformation in news articles.

International Fact-Checking Network. 2024. Fighting the Infodemic: The CoronaVirusFacts Alliance. https://www.poynter.org/coronavirusfactsalliance/. [Online; accessed 01-April-2024].

Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):630–638.

OpenScienceCollaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349:6251:943–6251:951.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Jacob N. Shapiro, Jan Oledan, and Samikshya Siwakoti. 2020. Fighting the Infodemic: The CoronaVirusFacts Alliance. https://esoc.princeton.edu/publications/esoc-covid-19-misinformation-dataset. [Online; accessed 01-April-2024].

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Simine Vazire. 2018. Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13:4:411—4:417.