An Empirical Assessment of the Qualitative Aspects of Misinformation in Health News

Chaoyuan Zuo, Qi Zhang, and **Ritwik Banerjee**
Computer Science, Stony Brook University
{chzuo, qizhang5, rbanerjee}@cs.stonybrook.edu

Abstract

The explosion of online health news articles runs the risk of the proliferation of low-quality information. Within the existing work on fact-checking, however, relatively little attention has been paid to medical news. We present a health news classification task to determine whether medical news articles satisfy a set of review criteria deemed important by medical experts and health care journalists. We present a dataset of 1,119 health news paired with systematic reviews. The review criteria consist of six elements that are essential to the accuracy of medical news. We then present experiments comparing the classical token-based approach with the more recent transformer-based models. Our results show that detecting qualitative lapses is a challenging task with direct ramifications in misinformation, but is an important direction to pursue beyond assigning *True* or *False* labels to short claims.

1 Introduction

In recent years, health information-seeking behavior (HISB) – which refers to the ways in which individuals seek information about their health, risks, illnesses, and health-protective behaviors (Lambert and Loiselle, 2007; Mills and Todorova, 2016) – has become increasingly reliant on Online news articles (Fox and Duggan, 2013; Medlock et al., 2015; Basch et al., 2018). Some studies also posit that with increasing involvement of the news media in health-related discussions, and direct-to-consumer campaigns by pharmaceutical companies, people are turning to the Internet as their first source of health information, instead of healthcare practitioners (Jacobs et al., 2017). This behavior is primarily driven by the users' need to gain knowledge (Griffin et al., 1999) about some form of intervention (e.g., drugs, nutrition, diagnostic and screening tests, dietary recommendations, psychotherapy). Furthermore, and perhaps counter-intuitively, information seekers seldom spend a lot of time on health

---

News headline: Experts warn coronavirus is 'as dangerous as Ebola' in shocking new study.
Source: www.express.co.uk/life-style/health/1275700/ebola-elderly-patients-coronavirus-experts-study-research-death-figures
Published: Apr 30, 2020 *Accessed*: March 21, 2021
Cause of misinformation
Comparing numbers from two different contexts: (1) the <u>hospital</u> fatality rate of COVID-19, and (2) the overall <u>case</u> <u>fatality rate</u> of Ebola.

Table 1: Medical misinformation due to a lack of understanding of domain-specific terminology.

websites. Instead, they repeatedly jump between search engine results and reading health-related articles (Pang et al., 2014, 2015).

In stark contrast to HISB, there is also growing lack of trust in the accuracy of health information provided on the Internet (Massey, 2016). This is perhaps to be expected, given how widespread health-related misinformation has become. For instance, in surveys where expert panels have judged the accuracy of health news articles, nearly half were found to be inaccurate (Moynihan et al., 2000; Yavchitz et al., 2012). Health-related misinformation, however, is rarely a binary distinction between *true* and *fake* news. In medical news, multiple aspects of an intervention are typically presented, and a loss of nuance or incomplete understanding of the process of medical research can lead to various types of qualitative failures, exacerbating misinformation in this domain.

Recently, news articles citing leading medical journals have suffered because of this. Table 1 shows an example that was disseminated widely in the United Kingdom, where technically correct facts were juxtaposed with misleading contexts – the *case* fatality rate of Ebola was incorrectly compared with the *hospital* fatality rate of COVID-19 (Winters et al., 2020). Indeed, medical misinformation is often a correct fact presented in an incorrect context (Southwell et al., 2019). Moreover, health-related articles are also known to present

76

(1)	Does the story/news release adequately discuss the costs of the intervention?
(2)	Does the story/news release adequately quantify the benefits of the intervention?
(3)	Does the story/news release adequately explain/quantify the harms of the intervention?
(4)	Does the story/news release seem to grasp the quality of the evidence?
(5)	Does the story/news release commit disease-mongering?
(6a)	Does the story use independent sources and identify conflicts of interest?
(6b)	Does the news release identify funding sources & disclose conflicts of interest?
(7)	Does the story/news release compare the new approach with existing alternatives?
(8)	Does the story/news release establish the availability of the treatment/test/product/procedure?
(9)	Does the story/news release establish the true novelty of the approach?
(10a)	Does the story appear to rely solely or largely on a news release?
(10b)	Does the news release include unjustifiable, sensational language, including in the quotes of researchers?

Table 2: **Review criteria.** The ten criteria for public relations news releases are almost identical to the ones for news stories (except for 6 and 10).

"disease-mongering", where a normal state is exaggerated and presented as a condition or a disease (Wolinsky, 2005).

Given how these issues are specific to medical misinformation, and how intricately the accuracy of medical facts is intertwined with the quality of health care journalism, the imperative to move beyond a binary classification of *true* and *fake* becomes clear. To this end, a set of specific principles and criteria have been proposed by scientists and journalists, based largely on the acclaimed work by Moynihan et al. (2000) and the Statement of Principles by the Association of Health Care Journalists (Association of Health Care Journalists, 2007).

We present a dataset (Sec. 2) specifically tailored for health news, and labeled according to a set of domain-specific criteria by a multi-disciplinary team of journalists and health care professionals. The detailed data annotation was carried out from 2006 to 2018 (Schwitzer, 2006). For each criterion, we present a classification task to determine whether or not a given news article satisfies it (Sec. 3), and discuss the results. Finally, we present relevant prior work (Sec. 4) before concluding.

2   Dataset

Our data is collected from Health News Review (Schwitzer, 2006)[1], which contains systematic re-

---

News headline: Virtual reality to help detect early risk of Alzheimer's
Source: www.theguardian.com/society/2018/dec/16/alzheimers-dementia-cure-virtual-reality-navigation-skills
Published: Dec 16, 2018 *Accessed*: April 26, 2021
Criterion labeled "not applicable":   (2) Does the story adequately quantify the benefits of the treatment/test/product/procedure?

Table 3: **Review criteria not applicable.** In this example, the study being reported has not yet taken place, so criterion (2) in Table 2 is not germane.

views of 2,616 news stories and 606 public relations (PR) news releases from a period of 13 years, from 2006 to 2018. Ten specific and standardized criteria were used for the reviews. These were chosen to align with the needs of readers seeking health information, and are shown in Table 2. The dataset consists only of articles that discuss a specific medical intervention, since the review criteria were deemed by journalists as being generally not applicable to discussions of multiple interventions or conditions. Each article is reviewed by two or three experts from journalism or medicine, and the results for each criterion include *Satisfactory*, *Not Satisfactory* and *Not Applicable*. The last label is reserved for cases where it is impossible or unreasonable for an article to address that criterion. Table 3 illustrates the utility of this label with one example from the dataset.

Going beyond the reviews themselves, we then collect the news articles being reviewed from the original news sites. However, nearly 30% of those pages have ceased to exist. Further, some articles could not be retrieved due to paywalls. Multiple prominent news organizations are featured in this data, with Fig. 1 showing the distribution over these organizations (for brevity, we show the top ten entities, with the tenth being "others").

Our final dataset comprises 1,119 articles (740 news stories and 379 PR news releases) along with their criteria-driven reviews. These are maintained as $(n, \{c_i\})$ tuples, where n is the news article, and c_i are the review results for each criteria. Since criteria 6 and 10 are slightly different for news stories and PR releases, we remove these from our empirical experiments. Further, we also remove criteria 5 and 9, since these require highly topic-specific medical knowledge. We do this in order to have our approach reflect the extent of medical knowledge available to the lay reader, who is unlikely to fully comprehend the specialized language of medical research publications (McCray, 2005).
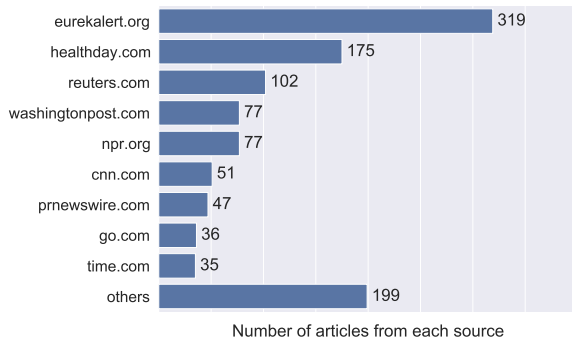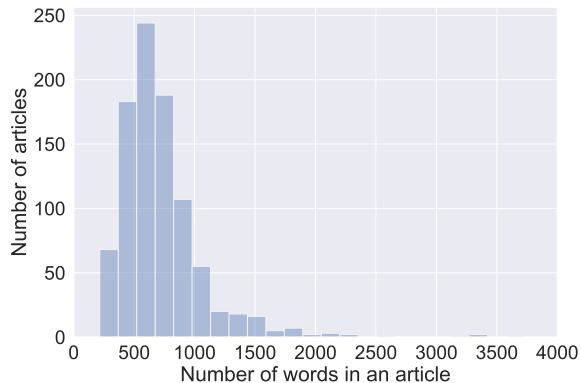
Figure 1: Distribution over news organizations.

Figure 2: The distribution of the size of news articles.

3 Experiments

We approach the problem as a series of supervised classification tasks, where the performance is evaluated separately for each review criterion. Moreover, since the reviewers use the *Not Applicable* label based on additional topic-specific medical knowledge, we discard the $(n, \{c_i\})$ tuples where c_i carries this label. This eliminates approximately 2.35% of the total number of tuples in our dataset, and paves the way for a binary classification task where each article is deemed satisfactory or not for the criterion c_i. The numbers of remaining news for each criterion are as shown in below Table 4.

In all experiments, we use 70% of the data for training. The rest is used as the test set. As a simple baseline, we use the Zero Rule (also called ZeroR or 0-R), which uses the base rate and classifies according to the prior, always predicting the majority class. We then experiment with the classical representation using TF-IDF feature encoding, as well as the state-of-the-art transformer-based models. In both approaches, we use 5-fold cross-validation during training to select the best hyperparameters for each model. These are described next.

| | *Number of news* | | | |
Criteria	Training	Test	Total	% of Positive samples
1	651	273	924	20.5
2	774	331	1105	30.9
3	733	309	1042	32.4
4	781	336	1117	34.3
7	745	316	1061	47.8
8	684	288	972	70.3

Table 4: **Data distribution across review criteria.** The percentage of positive samples for each criterion is shown in the last column. Note that the classes are quite imbalanced for every criteria except 7.

3.1 Models

For the feature-based models, we perform some preprocessing, which consists of removing punctuation, converting the tokens into lowercase, removing function words, and lemmatization. We use two supervised learning algorithms: support vector machines (SVM) and gradient boosting (GB). As noted in Table 4, our dataset suffers from class imbalance for every criteria except for one. Thus, for the remaining five criteria, we use adaptive synthetic sampling, viz., ADASYN (He et al., 2008). Further, to reduce the high dimensions of the feature space, we apply the recursive feature elimination algorithm from Scikit-learn (Buitinck et al., 2013) with SVM. In this process, the estimator is trained on the initial set of features, and the importance of each feature is determined by the weight coefficient. The least important features are then pruned. We recursively apply this process by selecting progressively smaller feature sets, until the 300 best features remain.

Next, we use several transformer-based models. Namely, BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), DistilBERT (Sanh et al., 2019) and Longformer (Beltagy et al., 2020). The maximum sequence length is set to 512 for every model, except for Longformer, for which the value is 4,096. We use random undersampling to mitigate the class imbalance, since the model's performance would otherwise be similar to the Zero Rule baseline.

3.2 Results and discussion

The results of our experiments are shown in Table 5. As the dataset is imbalanced for all but one criterion, our simple baseline is the Zero Rule instead of a random baseline. We measure the classifier

Criteria	0R F1	P	R	SVM F1	P	R	SVM*† F1	P	R	GB F1	P	R	GB* F1	P	R	GB*† F1	P	R
1	44.4	39.9	50.0	57.7	73.3	57.0	60.2	70.0	58.8	**66.3**	**76.3**	**63.6**	63.8	70.6	61.8	63.4	71.9	61.3
2	41.2	35.0	50.0	**64.4**	65.4	**63.8**	64.2	64.2	64.2	56.7	**65.9**	57.3	58.5	59.3	58.2	60.9	60.9	60.8
3	40.1	33.5	50.0	57.2	57.4	57.0	57.9	58.0	57.8	61.0	63.3	60.7	65.4	68.9	64.5	**67.4**	**67.9**	**67.0**
4	39.8	33.2	50.0	60.1	61.0	59.8	60.7	61.2	60.4	53.8	55.8	54.3	61.1	63.5	60.7	**68.1**	**69.1**	**67.5**
7	34.0	25.8	50.0	55.0	55.0	55.2	55.7	55.7	55.7	58.4	59.3	58.8	-	-	-	53.4	53.6	53.6
8	42.4	36.8	50.0	52.1	56.3	53.2	54.3	55.8	54.2	50.8	51.5	51.1	51.9	52.1	51.9	56.1	56.0	56.8

Criteria	BERT F1	P	R	ALBERT F1	P	R	XLNet F1	P	R	RoBERTa F1	P	R	DistilBERT F1	P	R	Longformer F1	P	R
1	55.6	55.4	56.4	57.2	59.1	63.9	63.6	62.8	65.8	62.6	62.7	68.4	62.8	63.0	69.1	62.8	62.2	66.4
2	48.5	55.7	56.1	52.0	52.4	52.7	60.9	60.7	62.0	60.8	61.0	62.7	54.6	56.4	57.6	62.2	62.3	64.2
3	58.2	58.1	58.3	55.4	55.5	55.5	55.9	56.1	56.6	59.4	59.3	60.0	53.1	57.9	54.4	63.8	63.8	65.4
4	40.7	58.2	50.2	49.8	54.0	54.0	47.6	57.0	56.1	56.1	58.2	59.2	55.6	57.0	57.9	50.2	52.4	52.7
7	61.2	62.8	62.1	40.2	58.1	52.0	48.4	55.9	53.4	**62.3**	62.7	62.6	57.4	57.5	57.4	56.8	60.0	58.8
8	54.7	57.6	59.7	55.8	56.1	55.6	54.6	54.7	55.3	**58.8**	61.1	**64.3**	55.5	57.6	59.9	53.5	58.7	54.3

Table 5: **Experiment results.** Models trained with oversampled data are marked with *. Models for which feature selection was performed are marked with †. For criterion 7 (see Tables 2 and 4), oversampling was not performed.

performances using macro-average of precision, recall, and F_1-score.

Gradient boosting achieves better performance on criteria 1, 3, 4, and 8. Also, the introduction of oversampling and feature selection increases the model performance for some criteria but not uniformly across the board.

The feature-based models outperform the transformer-based models in the first four criteria. We suspect this is mainly due to the size of the dataset after undersampling. We also check the number of words for the news collected (Fig 2), and more than half of which have more than 512 words. However, the Longformer model with maximum sequence length 4,096 does not achieve significantly better performance than other transformer-based models. The reason might be the "inverted pyramid" structure of news articles, which places essential information in the lead paragraph (Pöttker, 2003). We also notice that the first four criteria are more specific than the rest. For example, the first criterion is about the cost of the intervention, which could be answered by token-level searching. It is still a challenging task, however, given that even human readers find it difficult to follow the review criteria without expert training.

4 Related Work

For many years now, concerns have been raised about medical misinformation in the coverage by news media (Moynihan et al., 2000; Ioannidis, 2005). Moynihan et al. studied 207 news stories about the benefits and risks of three medications to prevent major diseases, and found that 40% of the news did not report benefits quantitatively while only 47% mentioned potential harms.

Various tasks and approaches have been formulated (Thorne and Vlachos, 2018) for fact-checking information. Multiple datasets have also been put forth. Ferreira and Vlachos (Ferreira and Vlachos, 2016) released a collection of 300 claims with corresponding news. This dataset was later significantly enlarged in the fake news challenge (Pomerleau and Rao, 2017). At a similarly large scale, Wang introduced a dataset comprising 12.8K manually labeled statements from POLITIFACT.COM and treated it as a text classification task. A large body of work, however, has dealt with fact-checking of short claims, both for fact-checking (Hassan et al., 2017) as well as for identifying what to check (Nakov et al., 2018).

Furthermore, a vast majority of prior work was on political news, while medical misinformation remained relatively neglected until its impact was underscored by the COVID-19 pandemic (e.g., Hossain et al. (2020); Serrano et al. (2020), among others). This body of work, however, continues to assign true/false labels or binary stance labels to short claims. In contrast, our work analyzes long articles and identifies whether or not they satisfy various qualitative criteria specifically important to medical news, as determined by journalists and health care professionals.

5 Conclusion

We present a first empirical analysis of qualitative reviews of medical news, since the traditional true/fake dichotomy does not adequately capture

the nuanced world of medical misinformation. To this end, we collect a dataset of medical news along with their detailed reviews based on multiple criteria. The novelty of this work lies in highlighting the importance of a deeper review and analysis of medical news to understand misinformation in this domain. For example, misinformation may easily be caused by the use of sensational language, or disease-mongering, or not disclosing a conflict of interest (all of which are criteria used in this work).

Our results show that this is a challenging task. The data reveals that for most of the criteria, less than half of the news articles are satisfactory. The commonly perceived notion of reputation notwithstanding, several articles from well-known sources (such as the ones shown in Fig. 1) also fall short of these qualitative benchmarks set by domain experts. This presents a clear data-driven picture of how the qualitative aspects of misinformation defy our expectations. We have presented a first step in this direction, and our hope is that this work leads to collaborative creation of similar datasets at larger scale by computer scientists and journalists, and in multiple domains even outside of health care.

Acknowledgment

References

Corey H. Basch, Sarah A. MacLean, Rachelle-Ann Romero, and Danna Ethan. 2018. Health Information Seeking Behavior Among College Students. Journal of Community Health, 43:1094–1099.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pages 108–122.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. ACL.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In Proc. of the 2016 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1163–1168.

Susannah Fox and Maeve Duggan. 2013. Health Online 2013. Internet & Technology, Pew Research Center. Last accessed: May 31, 2020.

Robert J. Griffin, Sharon Dunwoody, and Kurt Neuwirth. 1999. Proposed Model of the Relationship of Risk Information Seeking and Processing to the Development of Preventive Behaviours. Environmental Research, 80(2):S230–S245.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. Proc. of the VLDB Endowment, 10(12):1945–1948.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In Proc. of the IEEE Joint Conference on Neural Networks (IJCNN), 2008., pages 1322–1328. IEEE.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In Proc. of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics.

John P. A. Ioannidis. 2005. Why most published research findings are false. PLOS Medicine, 2(8):e124.

Wura Jacobs, Ann O. Amuta, and Kwon Chan Jeon. 2017. Health information seeking in the digital age: An analysis of health information seeking behavior among US adults. Cogent Social Sciences, 3(1):1302785.

Sylvie D. Lambert and Carmen G. Loiselle. 2007. Health Information-Seeking Behavior. Qualitative Health Research, 17:1006–1019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In 8th International Conference on Learning Representations ICLR 2020. OpenReview.net.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.

Association of Health Care Journalists. 2007. Statement of Principles of the Association of Health Care Journalists. https://healthjournalism.org/secondarypage-details.php?id=56. Last accessed: March 14, 2021.

Philip M. Massey. 2016. Where Do U.S. Adults Who Do Not Use the Internet Get Health Information? Examining Digital Health Information Disparities From 2008 to 2013. Journal of Health Communication, 21(1):118–124.

Alexa T. McCray. 2005. Promoting Health Literacy. Journal of the American Medical Informatics Association, 12(2):152–163.

Stephanie Medlock, Saeid Eslami, Marjan Askari, Derk L Arts, Danielle Sent, Sophia E de Rooij, and Ameen Abu-Hanna. 2015. Health information-seeking behavior of seniors who use the internet: A survey. J Med Internet Res, 17(1):e10.

Annette Mills and Nelly Todorova. 2016. An Integrated Perspective on Factors Influencing Online Health-Information Seeking Behaviours Information Seeking Behaviours. In ACIS 2016 Proc., page 83. Association for Information Systems.

Ray Moynihan, Lisa Bero, Dennis Ross-Degnan, David Henry, Kirby Lee, Judy Watkins, Connie Mah, and Stephen B. Soumerai. 2000. Coverage by the News Media of the Benefits and Risks of Medications. New England Journal of Medicine, 342(22):1645–1650.

Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In International Conference of the Cross-Language Evaluation Forum for European Languages, pages 372–387. Springer.

Patrick Cheong-Iao Pang, Shanton Chang, Jon M Pearce, and Karin Verspoor. 2014. Online Health Information seeking Behaviour: Understanding Different Search Approaches. In PACIS 2014 Proc., page 229. Association for Information Systems.

Patrick Cheong-Iao Pang, Karin Verspoor, Jon Pearce, and Shanton Chang. 2015. Better Health Explorer: Designing for Health Information Seekers. In Proc. of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, pages 588–597.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge. http://fakenewschallenge.org/.

Horst Pöttker. 2003. News and its communicative quality: The inverted pyramid – when and why did it appear? Journalism Studies, 4(4):501–511.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108.

Gary Schwitzer. 2006. Health News Review. https://www.healthnewsreview.org. Last accessed: March 22, 2021.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In Proc. of the 1st Workshop on NLP for COVID-19 at ACL 2020.

Brian G. Southwell, Jeff Niederdeppe, Joseph N. Cappella, Anna Gaysynsky, Dannielle E. Kelley, Emily B. Oh, April Peterson, and Wen-Ying Sylvia Chou. 2019. Misinformation as a Misunderstood Challenge to Public Health. American Journal of Preventive Medicine, 57.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In Proc. of the 27th International Conference on Computational Linguistics, pages 3346–3359. Association for Computational Linguistics.

William Y. Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Maike Winters, Ben Oppenheim, Jonas Pick, and Helena Nordenstedt. 2020. Creating misinformation: how a headline in The BMJ about covid-19 spread virally. BMJ, 369:m2384.

Howard Wolinsky. 2005. Disease mongering and drug marketing: Does the pharmaceutical industry manufacture diseases as well as drugs? EMBO reports, 6(7):612–614.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 5753–5763. Curran Associates, Inc.

A Yavchitz, I Boutron, A Bafeta, I Marroun, P Charles, J Mantz, and P Ravaud. 2012. Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. PLoS Medicine, 9(9):e1001308.