Evaluation of Applied Machine Learning for Health Misinformation Detection via Survey of Medical Professionals on Controversial Topics in Pediatrics

HAMMAN SAMUEL[*]

Department of Computing Science, University of Alberta, Edmonton, Canada

OSMAR ZAÏANE

Department of Computing Science, University of Alberta, Edmonton, Canada

FRANÇOIS BOLDUC

Department of Pediatrics, University of Alberta, Edmonton, Canada

In this research, we present an evaluation of a system for detection of health misinformation using applied machine learning. The system incorporates computing automation, information retrieval, and natural language processing in conjunction with evidence-based medicine to generate a veracity score based on consensus from trusted medical knowledge bases. For our study, we pre-computed the veracity scores of controversial topics in pediatrics with our proposed system, and then also solicited evaluations of these topics from medical professionals in the neurodevelopmental field via a quantitative survey. Hence, this work provides a double-blind comparison on the veracity of medical claims between our proposed system's results and medical professionals' responses. The results showed that our system's automated assessment matched professional opinions of medical personnel with 80% precision. The survey also demonstrated the inherent challenge with health misinformation detection, as there was no consensus among the medical professionals for 50% of the controversial statements. Nevertheless, this evaluation shows promising results for using objective trust metrics such as the veracity score, in contrast with subjective trust metrics that rely on potentially biased crowdsourcing, ratings, and pre-trained labelling of data.

CCS CONCEPTS • Computing Methodologies • Artificial Intelligence • Natural Language Processing • Information Extraction

Additional Keywords and Phrases: Health Misinformation, Applied Machine Learning, Social Media

ACM Reference Format:

Hamman Samuel, Osmar Zaïane, François Bolduc. 2021. Evaluation of Applied Machine Learning for Health Misinformation Detection via Survey of Medical Professionals on Controversial Topics in Pediatrics. In ICMHI '21: International Conference on Medical and Health Informatics, May 14–16, 2021, Kyoto, Japan. 8 pages.

1 INTRODUCTION

Not too long ago, viral social media posts were used to falsely associate vaccinations with autism [1]. Articles supposedly written by medical professionals that linked autism and vaccinations were heavily shared on Facebook and other social networks, leading to a perception among users that vaccinations are harmful.

[*] Correspondence email: hwsamuel@ualberta.ca

Needless to say, not getting vaccinated would give rise to more disease outbreaks and negatively affect public health overall. This is even more evident with the current COVID-19 pandemic, which itself has turned into an infodemic as social media discourse has been flooded with misinformation. In these situations, consensus-based methods relying on the "wisdom of the crowds", likes, or votes can be detrimental, and credible information from medical experts is needed.

Medical experts are able to determine trustworthiness of health information through Evidence-Based Medicine (EBM), a systematic approach for appraising health information on the basis of the best current evidence, clinical expertise, and patient needs in order to facilitate decisions about patient care [2]. Medical knowledge is health information verified through the scientific process and evidence. EBM arranges pertinent information into a hierarchy of evidence based on methodological quality. From the most reliable Level I up to Level VII, evidence can be grouped into systematic reviews of randomized controlled trials, well-designed randomized controlled trials, quasi-experimental studies, cohort studies, meta-synthesis, single qualitative studies, and reports of expert committees [3].

Computing automation can be applied in conjunction with EBM to determine the veracity of online health information. To this end, our proposed system algorithm was developed based on EBM and trusted medical information sources, in order to empower and educate online users to determine health information veracity. Our system is named MedFact[1] and addresses the challenges of layperson versus technical vocabularies, and issues of effectively presenting veracity of information in simplified and non-technical formats. Previously, our system has been evaluated using a survey of laypersons, as well as datasets with clear boundaries for false and true health information. In this paper, we extend our evaluation to ensure robustness and evolution of our system by evaluating it against a double-blinded survey of medical professionals in pediatrics.

2 SYSTEM OVERVIEW

Our proposed system is summarized as a three-step algorithm, involving forming a search query from incoming text with unknown veracity, followed by searching medical knowledge for related articles, and concluding with comparison between unknown claims and extracted medical facts.

2.1 Forming Search Query

In this stage, a supervised learning approach is used to build a binary classifier that labels a given phrase as medical or non-medical. The classifier is implemented as a Multi-Layer Perceptron (MLP) neural network, and medical phrases are input as word embeddings, with output of 0 if the phrase is non-medical or 1 if medical. In order to train the classifier, two categories of datasets were used. The first category corresponds to the "medical" label, including medical phrases from the SNOMED database and layperson health terms from the Consumer Health Vocabulary (CHV) dataset. SNOMED is a digital collection of medical terms provided by the U.S. National Library of Medicine [4]. The CHV dataset provides mappings of common layperson medical terms to technical terms in UMLS [5]. The second category corresponds to the "non-medical" label and contains known non-medical corpora from the Simple English Wikipedia (SEW) dataset [6]. From these datasets, a training sample is created by arbitrary selection of approximately 80% of the phrases from each dataset. A test sample of 20% is kept for internal scoring purposes. The phrases (hyphenated) are converted to word embeddings

[1] GitHub source code https://github.com/hwsamuel/MedFact

using the Word2vec model trained on medical corpora with skip-grams. The phrases and their corresponding labels from the training sample are used to train the MLP. The arbitrary selection process is repeated a number of times to achieve non-exhaustive cross-validation and the best trained model is used. The end result of this step is a set of medical keywords that are used as a search query in the next step.

2.2 Searching Medical Knowledge

Using the medical keywords query generated from the previous step, credible medical knowledge is searched via the TRIP database. TRIP focuses on evidence-based medical literature from various trusted sources including the U.S. National Library of Medicine (NLM) MEDLINE and PubMed articles, the Cochrane database of systematic reviews, the Database of Abstracts of Reviews of Effects (DARE), among others. Moreover, the TRIP database also searches within patient-friendly resources such as Cochrane Clinical Answers and WebMD's Medscape [7]. Results are categorized into the levels of evidence and can be sorted by quality, relevance, or date. A publication score is used to assess and rank quality of the results by incorporating the levels of evidence, Level I receiving the highest weight and subsequent levels receiving progressively lower weights. TRIP's quality metric is used to sort articles and incorporate strength of the evidence. Additional ranking of the articles is performed in order to evaluate the usefulness of the top-*n* articles based on their position in the results using Normalized Discounted Cumulative Gain (NDCG). After the articles are ranked, phrases are extracted from them via phrase chunking. Each chunked phrase extracted from the medical articles is compared with the search query keywords, and chunked phrases that do not correlate with search keywords are discarded because they will not be useful in the next steps.

2.3 Comparing Unknown Claims and Medical Facts

Given two phrases, one with unknown veracity and one with known veracity from a trusted medical source, the agreement between the phrases is determined using a shallow Convolutional Neural Network (CNN) architecture, which is more suitable for learning from smaller-sized labeled training datasets [8]. The shallow CNN incorporates semantic similarity and sentiment analysis of the two phrases. The feature set consists of the word embeddings of the two phrases, and sentiment information for each phrase, specifically polarity and subjectivity [9]. Also, the negation modifier is used from dependency parsing [10] of the related sentence containing the target phrases as an additional binary feature, where 1 implies the presence of the negation modifier and 0 means an absence. The training dataset was built from Medical Science Stack Exchange, an online question-answering community where users can post health-related questions, and moderators manually flag semantically equivalent posts as duplicates. The training dataset consists of pairs of phrases extracted from the duplicate posts' title and body using phrase chunking. Ultimately, given two phrases, the veracity score is defined using the shallow CNN classifier's output label’s associated probability, and the overall veracity score for a paragraph is determined by averaging the veracity scores of its constituent phrases.

3 RELATED LITERATURE

Research on trust in social media falls into two categories: empirical analysis and algorithmic contributions. Various studies have been conducted to measure the usefulness of generic trust metrics in online communities. These empirical studies can further be grouped into three categories looking at either the network structure, content, or behavioral signals from users. The network structure and its properties help to iteratively determine trust of a given user based on relationships to other trusted users [11,12]. Content has also been investigated as an indicator for trustworthiness. However, content assessment in current approaches relies on reputation assessment which is limited by user-based ratings. Collaborative content-based methods have also been investigated for determining user reputation [13]. Other metrics such as frequency and sentiment of follow-up posts in relation to an original post have also been studied.

Research on pragmatic contributions to trust in health information were fewer until the COVID-19 pandemic. The seminal work by [14] on HealthTrust was one of the earlier health information-focused studies on trust. HealthTrust automatically assesses new health information based on a set of health web sites with known credibility. Comparison is based on link analysis and content-based analysis. In link analysis, the assumption is that trustworthy content will point to trustworthy web sites as an appeal for authority. Consequently, TrustRank is used to infer a ranking for new content based on inbound and outbound link analysis. In content-based analysis, topic discovery via the TAGME algorithm [15] is used to classify new content as suspicious or trustworthy based on topic similarity with known content via affinity propagation clustering. Secondly, to improve content matching, Hidden Markov Models are applied to an annotated training set in order to model trustworthy and suspicious sentences. A HealthTrust score is assigned for each web site, which is then iteratively exploited.

Recently, there have been many works published in preprint focusing on detection of health misinformation related to COVID-19. The majority of these methodologies can be grouped as either semi-supervised or supervised machine learning. These methods require annotated training data to identify misinformation [16,17]. To support this methodology, various datasets have been annotated independently as well as from fact-checking websites and fact-checked articles covering a broad range of political and medical topics [18,19]. Veracity of specific health topics such as cancer treatments has also investigated using machine learning techniques such as the study by [20]. Using a bag of words representation as the feature set, web pages with medical advice were labeled as positive or negative based on whether they contained questionable content, and the trained model used to assign new labels to new web pages. This approach relied on keyword co-occurrences and correlations instead of cross-referencing trusted medical knowledge.

4 METHODOLOGY

This survey was conducted as part of the ethics approval from the Research Ethics Board of the authors' institution. The survey provided a double-blind comparison on the veracity of medical claims between our system's results and medical professionals' responses. Hence, participants were not shown the results of our system, but rather were asked to independently evaluate statements related to pediatrics. Also, our system's computations for the same statements on pediatrics were computed prior to administering the survey.

A questionnaire was disseminated privately among known medical professionals from in the neurodevelopmental field in pediatrics to avoid layperson opinions. Six statements related to pediatrics were shown to the participant to rate each statement based on their professional evaluation of the statement's veracity using a psychometric scale: *Strongly Disagree*, *Disagree*, *Neutral*, *Agree*, *Strongly Agree*, and *Do Not Know*.

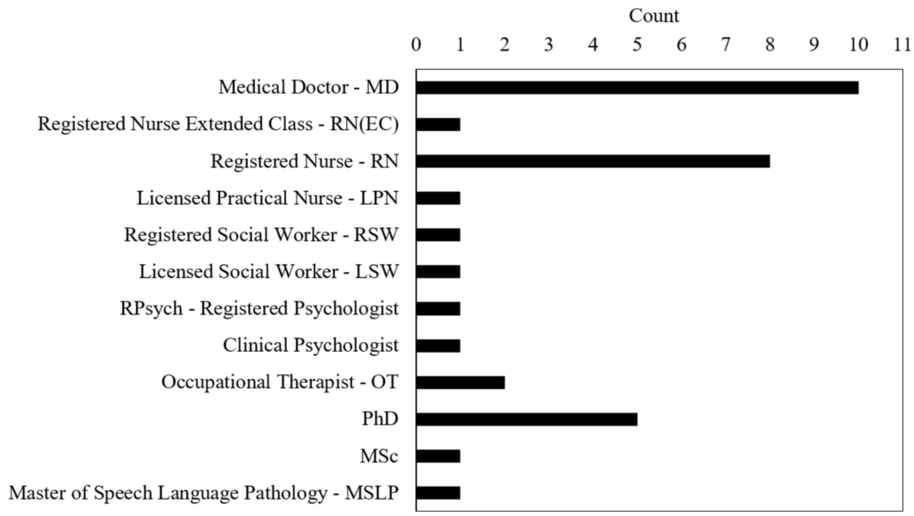
Each of the statements, selected from Facebook, Wikipedia, blogs, and news articles, belonged to one of the following topics: general pediatrics, autism, behavior, Applied Behavior Analysis (ABA), Attention Deficit Hyperactivity Disorder (ADHD), or Positive Parenting Program (PPP). For each participant, the six statements were selected from three rubrics, A, B, and C, and the statements within the selected rubric were then randomly re-ordered. Hence, each subsequent participant viewed a different set of six statements from each rubric, with the rubric selection being rotated in sequence. A total of 10 respondents viewed rubric A, 11 respondents were shown rubric B, and 13 viewed rubric C. The list of statements and rubrics used to administer the survey are detailed in Table 1.

Table 1: Survey Statements by Rubric and Topic

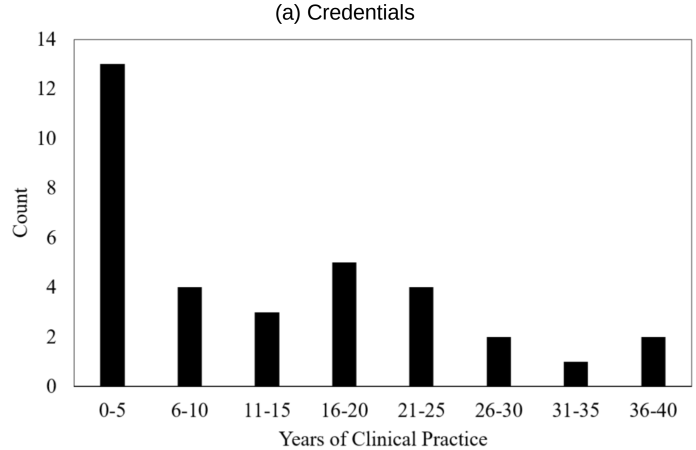
Rubric	ID	Statement	Topic
A	A1	A lot of government-published studies show vaccines cause autism	Autism
	A2	When dealing with a misbehaving child, intentionally ignore a problem behavior instead of reacting or giving negative attention to the child	Behavior
	A3	ABA therapy accounts for 45% of pediatric therapies that develop long-lasting and observable results	ABA
	A4	Parents of children with disabilities should not be allowed to use growth attenuation therapy	General
	A5	When ADHD is undiagnosed and untreated, ADHD contributes to problems succeeding in school and graduating	ADHD
	A6	A review of 33 studies published in BMC Medicine found no convincing evidence that Triple P interventions work across the whole population, or that any benefits are long-term	PPP
B	B1	Parents can change from using ineffective and coercive discipline such as physical punishment, shouting, and threatening to using effective strategies in specific situations	Behavior
	B2	Applied Behavioral Analysis (ABA) is based on a cruel premise - of trying to make people with autism “normal”	ABA
	B3	Homeopathic treatments for hyperactive children have been generally successful	ADHD
	B4	The age threshold for using medical intervention for children with gender dysphoria should be lowered	General
	B5	Environmental factors that could trigger predisposed genes to mutate and cause autism are vast and could include certain drugs, extensive television viewing, or infections during pregnancy	Autism
	B6	Triple P trials are particularly susceptible to risks of bias and investigator manipulation of apparent results	PPP
C	C1	Most scientists agree that genes are one of the risk factors that can make a child more likely to develop autism	Autism
	C2	The most serious problem with the Triple P literature is the over-reliance on positive but substantially underpowered trials	PPP
	C3	Selective Serotonin Reuptake Inhibitors (SSRIs) are an effective treatment for pediatric OCD	Behavior
	C4	A child with ADHD is accident-prone, likely to make careless mistakes, and take unnecessary risks	ADHD
	C5	Neurodiversity should be accepted as naturally different rather than abnormal and needing to be fixed	General
	C6	ABA is just animal training adapted for use with people	ABA

5 RESULTS AND DISCUSSION

The six statements in each rubric were from varied topics in pediatrics, and a total of 34 participants responded. Aggregated self-reported credentials, years of clinical practice, and areas of practice are shown in Figure 1. The highest years of clinical practice were 36, with mean of 13.44 years and median of 10.50 years.



(a) Credentials



(b) Years of Clinical Practice

Figure 1: Demographic Information of Participant Medical Professionals

The statements were evaluated by our system and a veracity score computed. Based on the score and confidence, a *System Label* was assigned to each statement. For comparison, the responses of the medical professionals were categorized as either in agreement, disagreement, or uncertain about each of the statements. Based on the majority consensus, a *Medic Label* was assigned to each statement. Ultimately, the two labels were compared to evaluate our system's corroboration with medical professionals, with details provided in Table 2.

Table 2: Comparison of Responses by Medical Professionals versus Proposed System

ID	System Veracity Score	System Label	Medics Label	Medics Disagree	Medics Uncertain	Medics Agree	Consensus among Medics
A1	0.11	**Untrusted**	**Untrusted**	0.90	0.10	0.00	Disagree
A2	0.67	Unknown	Trusted	0.30	0.20	0.50	Agree
A3	0.73	Trusted	Unknown	0.20	0.50	0.30	No consensus
A4	0.19	Untrusted	Unknown	0.40	0.60	0.00	No consensus
A5	0.78	**Trusted**	**Trusted**	0.00	0.10	0.90	Agree
A6	0.69	**Unknown**	**Unknown**	0.20	0.50	0.30	No consensus
B1	0.77	**Trusted**	**Trusted**	0.27	0.18	0.55	Agree
B2	0.66	Unknown	Untrusted	0.55	0.45	0.00	Disagree
B3	0.13	**Untrusted**	**Untrusted**	0.55	0.45	0.00	Disagree
B4	0.12	Untrusted	Unknown	0.27	0.55	0.18	No consensus
B5	0.80	Trusted	Unknown	0.36	0.45	0.18	No consensus
B6	0.61	**Unknown**	**Unknown**	0.27	0.64	0.09	No consensus
C1	0.69	**Trusted**	**Trusted**	0.00	0.08	0.92	Agree
C2	0.66	Unknown	Trusted	0.00	0.46	0.54	Agree
C3	0.04	Untrusted	Unknown	0.31	0.54	0.15	No consensus
C4	0.82	**Trusted**	**Trusted**	0.23	0.15	0.62	Agree
C5	0.47	**Unknown**	**Unknown**	0.08	0.54	0.38	No consensus
C6	0.11	Untrusted	Unknown	0.31	0.54	0.15	No consensus

It should be noted that the “Consensus among Medics” column denotes the overall opinion of the medical professionals in relation to the statement specified in the ID column, and its label computed based on the majority percentage of medics disagreeing, uncertain, or agreeing. The “Medics Label” column is accordingly set based on the consensus. When taking into consideration all the statements, our proposed system’s automated assessment matched the professional opinions of medical personnel by 50%. Even among the professionals, there was no consensus for 50% of the statements, and the statements were marked as uncertain, demonstrating the challenge with determining veracity, given the variety of topics. Excluding statements where professionals were uncertain, our system corroborated even closer with medical professionals. Focusing only on the statements that had agreement or disagreement among the medical professionals, and taking these as ground truth, the accuracy and recall of our system was 67%, with precision at 80%, and F1 score was 73%.

6 CONCLUSION

This research work provided details on the implementation and usability testing details of our proposed system and the veracity score as an objective trust metric. The usefulness of our proposed system was tested with a survey of medical professionals via a double-blind comparison on the veracity of medical claims between our proposed system’s results and medical professionals' responses. The results showed that our system’s automated assessment matched professional opinions of medical personnel with 80% precision. Our study also discussed and appraised the inherent challenge with health misinformation detection when there is no consensus among medical professionals for controversial statements. This evaluation shows promising results for using objective trust metrics such as the veracity score, in contrast with subjective trust metrics that rely on potentially biased crowdsourcing, ratings, and pre-trained labelling of data.

ACKNOWLEDGMENTS

We wish to thank the Alberta Machine Intelligence Institute (AMII) for funding this research work. We also thank the Kids Brain Health Network (KBHN) for their assistance in the dissemination of our survey, as well as the Women and Children Health Research Institute (WCHRI) for hosting the online survey via REDCap. We also acknowledge the help of Dr. Carrie Demmans Epp and Dr. Deena Hamza in formulating the questionnaire.

REFERENCES

[1] Anna Kata. 2012. Anti-Vaccine Activists, Web 2.0, and the Postmodern Paradigm–An Overview of Tactics and Tropes Used Online by the Anti-Vaccination Movement. *Vaccine* 30, 25 (2012), 3778–3789.

[2] Trisha Greenhalgh. 2010. *How to Read a Paper: The Basics of Evidence-Based Medicine*. John Wiley & Sons.

[3] Betty J Ackley. 2008. *Evidence-Based Nursing Care Guidelines: Medical-Surgical Interventions*. Elsevier Health Sciences.

[4] Ronald Cornet and Nicolette de Keizer. 2008. Forty Years of SNOMED: A Literature Review. *BMC Med. Inform. Decis. Mak.* 8, Suppl 1 (2008), S2.

[5] Catherine Smith and P Stavri. 2005. Consumer Health Vocabulary. *Consum. Heal. Informatics* (2005), 122–128.

[6] Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio. 2015. Extracting Knowledge from Text with PIKES. In *International Semantic Web Conference*.

[7] Jon Brassey. 2005. TRIP Database: Identifying High Quality Medical Literature from a Range of Sources. *New Rev. Inf. Netw.* 11, 2 (2005), 229–234.

[8] Rie Johnson and Tong Zhang. 2015. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*.

[9] Bo Pang, Lillian Lee, and others. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends® Inf. Retr.* 2, 1–2 (2008), 1–135.

[10] Marie-Catherine De Marneffe and Christopher D Manning. 2008. *Stanford Typed Dependencies Manual*.

[11] Iraklis Varlamis, Magdalini Eirinaki, and Malamati Louta. 2010. A Study on Social Network Metrics and their Application in Trust Networks. In *Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 168–175.

[12] Sonja Grabner-Kräuter and Sofie Bitter. 2015. Trust in Online Social Networks: A Multifaceted Perspective. *Forum Soc. Econ.* 44, 1 (2015), 48–68.

[13] Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2015. Collaborative Content-Based Method for Estimating User Reputation in Online Forums. In *Web Information Systems Engineering*. Springer, 292–299.

[14] Meeyoung Park. 2013. HealthTrust: Assessing the Trustworthiness of Healthcare Information on the Internet. University of Kansas.

[15] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-The-Fly Annotation of Short Text Fragments (By Wikipedia Entities). In *ACM International Conference on Information and Knowledge Management (CIKM)*, 1625–1628.

[16] Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid–A Multilingual Cross-domain Fact Check News Dataset for COVID-19. *arXiv Prepr. arXiv2006.11343* (2020).

[17] Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. 2020. No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection. *arXiv Prepr. arXiv2010.06906* (2020).

[18] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K Funk, Rodney Michael Kinney, Ziyang Liu, W Merrill, P Mooney, D Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B Stilson, A Wade, K Wang, Christopher Wilhelm, Boya Xie, D Raymond, Daniel S Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *arXiv 2004.05125* (2020).

[19] Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2020. Drink Bleach or Do What Now? Covid-HeRA: A Dataset for Risk-Informed Health Decision Making in the Presence of COVID19 Misinformation. *arXiv Prepr. arXiv2010.08743* (2020).

[20] Yin Aphinyanaphongs, Constantin Aliferis, and others. 2007. Text Categorization Models for Identifying Unproven Cancer Treatments on the Web. In *World Congress on Medical Informatics (MedInfo)*, 968.