MedFact: Towards Improving Veracity of Medical Information in Social Media using Applied Machine Learning

Hamman Samuel[0000-0002-3053-6047] and Osmar Zaïane[0000-0002-0060-5988]

Department of Computing Science, University of Alberta, Edmonton, Canada
hwsamuel@ualberta.ca, zaiane@ualberta.ca

Abstract. Since the advent of Web 2.0 and social media, anyone with an Internet connection can create content online, even if it is uncertain or fake information, which has attracted significant attention recently. In this study, we address the challenge of uncertain online health information by automating systematic approaches borrowed from evidence-based medicine. Our proposed algorithm, MedFact, enables recommendation of trusted medical information within health-related social media discussions and empowers online users to make informed decisions about the credibility of online health information. MedFact automatically extracts relevant keywords from online discussions and queries trusted medical literature with the aim of embedding related factual information into the discussion. Our retrieval model takes into account layperson terminology and hierarchy of evidence. Consequently, MedFact is a departure from current consensus-based approaches for determining credibility using "wisdom of the crowd", binary "Like" votes and ratings, popular in social media. Moving away from subjective metrics, MedFact introduces objective metrics. We also present preliminary work towards a granular veracity score by using supervised machine learning to compare statements within uncertain social media text and trusted medical text. We evaluate our proposed algorithm on various data sets from existing health social media involving both patient and medic discussions, with promising results and suggestions for ongoing improvements and future research.

1 Introduction

Fake news on social media has garnered considerable attention recently. Our research looks at a related problem in the medical domain where consumption of inaccurate and uncertain medical information can have life-threatening consequences. For example, viral social media posts were recently 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 many users that vaccinations are harmful. On the other hand, not getting vaccinated would give rise to more disease outbreaks and negatively affect public health overall. With the vast amount of information available online, certain information-seeking skill sets are needed to locate credible information, especially for sensitive topics like health information. Online content about personal well-being, management of diseases, and other medical topics related to medicine and health care is referred to as health information [2]. In contrast, medical knowledge is health information verified through the scientific process and Evidence-Based Medicine (EBM).

Medical experts are able to determine trustworthiness of health information through 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 [3]. 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 [4].

This study explores how computing automation can be applied in conjunction with EBM to determine the credibility of online health information. We develop MedFact, an algorithm based on EBM and trusted medical information sources, in order to empower and educate online users to determine the credibility of health information. We also address the challenges of layperson versus technical vocabularies, and issues of effectively presenting credibility of information in simplified and non-technical formats. We also present our solution to the research question of granular phrase-level textual agreement. As a side effect of our proposed approach, various aspects of the EBM methodology are automated, including information retrieval and processing. We use the terms “credibility” and “veracity” interchangeably as referring to factual accuracy of information. These concepts are closely related to the notion of “trust”, involving a willing interaction between two or more entities with an implicit belief that the interaction will at least be self-beneficial in the worst case, and mutually beneficial to all entities involved in the best case [5].

2 Background

2.1 Current State of Computational Research in Trust and Health Social Media

Research into credibility in social media falls into two categories of empirical analysis and algorithmic contributions. Various studies have been conducted to measure the effectiveness of generic trust metrics in forums and 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 [6]. 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 been proposed to determine user reputation. Other metrics such as frequency and sentiment of follow-up posts in relation to an original post have also been studied [7].

The popular approach for representing veracity is via ratings. There are various implicit and explicit metrics for trust requiring users to provide subjective feedback. Trust metrics provide an abstracted evaluation of the level of credibility or trust associated with content or users. Common trust metrics found in social networks are scaled unary ratings, such as Facebook “Like” , binary ratings such as up or down votes, ranked ratings such as Likert scale rankings, and reputation systems for measuring user trust using achievement levels, badges, and gamification [8]. Some drawbacks of ratings-based systems include inflation, bootstrapping, whitewashing, and cold start [9].

Research on pragmatic contributions to health information veracity are fewer. The seminal work by [10] on HealthTrust is one of the few 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 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 (HMM) are applied to an annotated training set in order to model trustworthy and suspicious sentences. A HealthTrust score is then assigned for each web site, which could then be iteratively exploited. However, there are no data sets available for use in training supervised learning models.

Veracity of specific health topics such as cancer treatments has also investigated using machine learning techniques such as the study by [11]. 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.

2.2 Psychological Viewpoint on Popularity of Fake Online Health Information

Various factors contribute to the present proliferation of unsafe health information online, which need to be taken into consideration when developing any approach for promoting credible information and preventing unsafe viral health campaigns. Apart from the development of technical solutions and effective trust metrics, the psychological biases of users consuming health information also need to be understood, including users' preference for layperson health stories, perceived resistance to medical facts, and the perception of medical expertise among laypersons.

Neural Coupling The information seeking behavior of laypersons and patients is based on story-telling rather than systematic medical and scientific methodologies. Patients tend to use personal experience and stories as a source of authoritativeness rather than scientific methodology [12]. This behavior is related to neural coupling, an effect observed in neuroscience between storytellers and listeners. Experiments have shown that when a storyteller is communicating with listeners, the listener's brain patterns will eventually mirror the storyteller's patterns. Neural coupling is an evolutionary trait to help human species to learn from each other through emotions [13]. The popularity of story-based narratives on health social media could also be attributed to these primal triggers. In the case of the "anti-vaxxers", even fake personal stories were effective in convincing people not to vaccinate because of the emotional format of the message [1].

Backfire Effect Studies related to anti-vaxxers attempted to investigate the effectiveness of counter-messages promoting vaccinations for Measles Mumps Rubella (MMR) [14]. In the study, anti-vaxxer parents of children needing MMR vaccinations were presented with various interventions. Firstly, they were presented with information on the lack of evidence associating autism to vaccinations. Secondly, they were shown textual information on risks of not getting vaccinated. Thirdly, images of other children who had contracted MMR-related diseases were shown. And finally, parents were told a dramatic story of a child who did not get vaccinated for measles and almost died. Surprisingly, none of the interventions were statistically significant in convincing the parents. In some cases, the parents' belief that vaccinations are harmful was even strengthened, for instance when being shown the imagery of sick children who did not get vaccinated. These counterintuitive results could be explained by the backfire effect, wherein the presentation of contradictory evidence is not only ineffective in convincing people, but leads people to strengthen their belief [15]. Related to the backfire effect is confirmation bias, where users online tend to seek out and gravitate towards information supporting their beliefs and ignore opposing viewpoints [16].

Dunning-Kruger Effect The Dunning-Kruger effect is attributed to unskilled persons having the illusion of superior competence [17], a trait that can be readily observed in the online health information communities, where laypersons eagerly and confidently provide medical advice to other laypersons. This phenomenon is clearer in the study of agnotology, where inaccurate or misleading scientific information is willfully promoted to induce ignorance about facts [18]. Essentially, online health information is saturated with information that is not credible, yet is being propagated due to users' willingness to look for quick solutions to complex health problems, such as autism [19].

3 Methodology

We define the task of determining the credibility of medical content as a five-step process. Given any textual document, such as a social media post, the first step is to extract health-related phrases $\{x_1, x_2, ..., x_m \in X\}$. The veracity of these phrases is unknown. The second step uses automated information retrieval and processing to search trusted scientific and medical knowledge bases for each of the unknown phrases $x_i \in X$. In this step, each trusted source would yield zero or more relevant articles, providing a collection of trusted articles which are ranked and filtered by relevance. Moreover, each trusted article would have various related credible phrases that are identified in the third step to generate a collection of trusted phrases $\{t_1, t_2, ..., t_n \in T\}$. The semantic similarity between a given trusted phrase, $t_j \in T$, and x_i is used for inferring an *agreement score*, $\Upsilon(x_i, t_j)$ between the two phrases. In the fourth step, an aggregated agreement score for a given unknown phrase is computed by comparing it with all trusted phrases and averaging the agreement score as formulated in Equation 1. In the fifth step, an overall *veracity score* ϑ is computed for the social media post from the aggregated agreement scores of all unknown phrases as shown in Equation 2.

$$\Upsilon(x_i) = \left(\sum_{p=1}^{n} \Upsilon(x_i, t_p) \right) / n \tag{1}$$

$$\vartheta = \left(\sum_{q=1}^{m} \Upsilon(x_q) \right) / m \tag{2}$$

Our methodology has parallels with the EBM five-step model: ask, acquire, appraise, apply, and analyze [20]. Overlapping MedFact with EBM, asking a question entails seeking to investigate the veracity of a social media post, while acquiring involves computationally gathering the available evidence related to the question. The overall pipeline for MedFact is depicted in Figure 1.

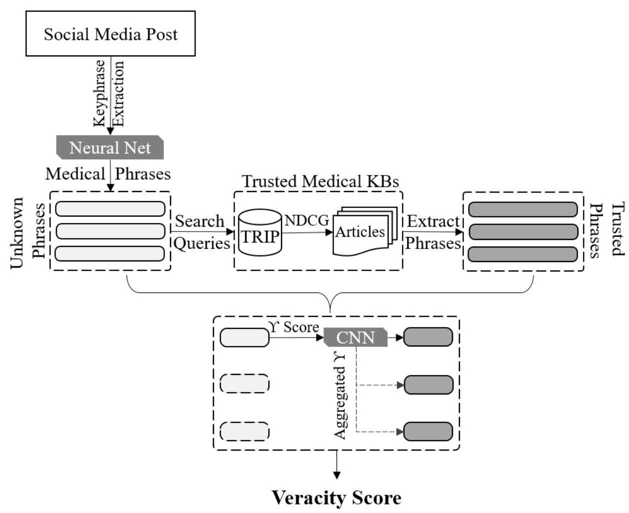


Fig. 1: Overview of MedFact Algorithm

Step 1 To extract relevant health phrases from a given social media posting, candidate phrases are extracted using key phrase extraction. The next stage identifies health-related phrases from among the candidate phrases. Extraction of key phrases is done using the TextRank algorithm[1]. In the next stage, we use a supervised learning approach to build a binary classifier that for classifying a given phrase as medical or non-medical. The classifier is implemented as an artificial 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 our classifier, we use two categories of data sets. The first category corresponds to the "medical" label, including medical phrases from the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) database and layperson health terms from the Consumer Health Vocabulary (CHV) data set.

[1] The GenSim Python API includes the TextRank algorithm [21] implementation https://radimrehurek.com/gensim/summarization/keywords.html

SNOMED CT[2] is a digital collection of medical terms provided by the U.S. National Library of Medicine [22]. The CHV data set[3] provides mappings of common layperson medical terms to technical terms in the Unified Medical Language System (UMLS) [23]. The second category corresponds to the "non-medical" label and contains known non-medical corpora from the Simple English Wikipedia (SEW) data set[4] [24]. From these data sets, a training sample is created by arbitrary selection of approximately 80% of the phrases from each data set. A test sample of 20% is kept for evaluation purposes. The phrases (hyphenated) are converted to word embeddings using the Word2Vec deep neural network model trained on medical corpora with skip-grams [25]. The phrases and their corresponding labels from the training sample are used to train our neural net. The arbitrary selection process is repeated a number of times to achieve non-exhaustive cross-validation and the best trained model is used.

Step 2 Credible medical knowledge can be queried from the Turning Research Into Practice (TRIP) database[5]. TRIP focuses on evidence-based medical literature from various trusted sources including the NLM's 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 [26]. 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. We use TRIP's quality metric to sort articles and incorporate strength of the evidence. We perform additional ranking of the articles in order to evaluate the usefulness of the top-n articles based on their position in the results using Normalized Discounted Cumulative Gain (NDCG) [27].

Step 3 In order to compare unknown phrases with trusted phrases, phrases are extracted from the ranked medical articles via phrase chunking. Firstly, each article's text is split using sentence and word tokenization. Next, Part-Of-Speech (POS) tagging is performed on the tokens, followed by phrase chunking[6] which segments the sentences into noun phrases. After that, each chunked phrase extracted from the medical articles is compared with the set of unknown phrases, and trusted phrases that do not correlate with unknown phrases are discarded because they will not be useful in the next steps.

---

[2] SNOMED CT data set available from U.S. National Library of Medicine (NLM) `https://nlm.nih.gov/healthit/snomedct`

[3] CHV data set available from the Consumer Health Vocabulary Initiative `http://consumerhealthvocab.org`

[4] SEW historical data set available via PIKES home page `http://pikes.fbk.eu/eval-sew.html`

[5] The TRIP database is accessible programmatically via web services that were most kindly made available to the authors by Jon Brassey, the TRIP database creator `https://tripdatabase.com/addtrip`

[6] POS tagging is done using the Penn Treebank tags set, all steps in this particular pipeline are programmed with the NLTK Python library `http://nltk.org`

Step 4 Given a phrase whose credibility needs to be ascertained, a corresponding set of phrases from a trusted source can be used as evidence for supporting or rejecting the unknown phrase as credible. We model this problem as that of predicting a class label over a pair of phrases, where two binary labels are possible: *Yes* and *No*. The former label implies that the two phrases have the same meaning, while the latter label means the phrases could contain incompatible propositions such as contradictions.

Given two phrases, we determine their agreement using deep learning, incorporating semantic similarity and sentiment analysis of the two phrases. Our feature set consists of the word embeddings of the two phrases, and sentiment information[7] for each phrase, specifically polarity and subjectivity [28]. Polarity for a phrase is in the range [-1.0, 1.0] where -1.0 implies very negative sentiment and 1.0 means very positive sentiment, while subjectivity values are in the range [0.0, 1.0] where 0.0 means very objective and 1.0 implies very subjective. We also use the negation modifier from dependency parsing [29] 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[8]. For our deep learning neural network implementation, we use a shallow Convolutional Neural Network (CNN) architecture[9], which is more suitable for learning from smaller-sized labeled training data sets [30]. We build our training data set from Health Stack Exchange (HSE)[10], an online question-answering community where users can post health-related questions[11]. Within the HSE community, moderators can manually flag semantically equivalent posts as *Duplicate*.

Our training data set consists of pairs of phrases extracted from the duplicate posts' title and body using phrase chunking. The related medical phrase pairs extracted from these question pairs are assigned the *Yes* label. For question pairs that are not duplicates, the *No* label is assigned to the related phrase pairs in the training data set. We subsequently manually curate the training data set for accuracy of the initial labeling in order to verify whether the phrase pairs are in agreement or not. Ultimately, given two phrases, the *agreement score* is defined using the classifier's output label. If the *No* label is returned, agreement score is 0, while for the *Yes* label, the score is 1.

Step 5 The veracity score enables aggregation of the agreement scores of many pairs of unknown phrases and their respective trusted phrases, and provides a single metric for measuring the credibility of a given social media posting or document. This approach allows for a granular definition of veracity starting from phrase-level agreement to document-level aggregated agreement. Depending on the number of unknown and trusted phrase pairs, the overall veracity score is computed as an average, hence it is within the range [0.0, 1.0], and can be expressed as a simplistic percentage value.

---

[7] Sentiment analysis is performed using the TextBlob Python library `http://textblob.readthedocs.io`

[8] The spaCy Python library is used for generating dependency trees `https://spacy.io`

[9] We implement a shallow CNN with the ConText tool `https://github.com/riejohnson/ConText`

[10] Health Stack Exchange's beta web site `https://health.stackexchange.com`

[11] Data set curated from the Stack Exchange Data Dump from the Internet Archive `https://archive.org/details/stackexchange`

4 Evaluation

4.1 Effectiveness of Key Phrase Extraction

We evaluated the performance of the key phrase extraction step using the HSE data set, which contains human-annotated tags per question. We compared our extracted key phrases with the annotated tags and used the recall metric to measure performance. For instance, if all the tags were found within the key phrases of a given question, the recall was recorded as 100%. The HSE data set contained 3,958 questions and 2,260 tag sets and the average recall for the key phrase extraction step was 81.28%.

4.2 Effectiveness of Medical Phrases Extraction

We used the SNOMED CT, CHV, and SEW databases to perform extraction of medical phrases from social media postings using a neural network for binary classification. We initially tested the performance of the classifier trained on combinations of medical and English data sets, and then evaluated the overall performance of the classifier using all three data sets, which provided the best precision and recall values of 74.00% and 67.10% respectively via 10-fold cross-validation.

4.3 Curation of Phrase Pairs for CNN Training

Using the best performance configuration for the medical key phrases extraction step, a total of 11,517 relevant medical phrases was extracted, averaging 2.91 phrases per HSE question. A total of 43 duplicate questions was recorded by the HSE moderators, and 175 phrases were extracted from these duplicates, out of which 83 pairs had agreement. We also manually inspected the rest of the data set to retrieve a total of 181 pairs that could be marked with the *Yes* label. The training data set was then balanced to arbitrarily include an equivalent number of phrase pairs with the *No* label. Overall, the average Fleiss Kappa for the curation process was 0.691, indicating “moderate agreement” for the corrections and additions made [31].

4.4 Appraisal of Phrase Agreement CNN

We explored the relationship between performance of the CNN used for determining phrase agreement and the feature set. The best precision and recall values of 0.606 and 0.448 respectively were achieved by including all features.

4.5 Feedback on Usage of Veracity Score

To assess the effectiveness of MedFact, we designed a short survey that was administered to 19 users. The survey contained polarizing social media postings on the link between vaccination and autism, apricot pits as a cure for cancer, and usefulness of flossing for dental care. Firstly, a posting supporting vaccination and autism was displayed, followed by a post debunking the notion. Similarly, users were then shown a posting supporting apricot pits as a cure for cancer, and then shown an opposing post.

Lastly, posts supporting and opposing the need to floss were shown. For each posting shown, the veracity score expressed as a percentage (rounded-off) was visible. The top 3 trusted articles related to the posting were also displayed. After displaying each posting, users were asked three questions related to the veracity score and the linked articles. Each question required a Yes or No response for the related post being displayed. Firstly, they were asked "*Is the veracity score useful in this context?*". Next, they were asked "*Is the veracity score accurate for this post?*". Lastly, they were asked "*Are the links to the medical articles useful?*". A summary of users' responses recorded for the questions is shown in Figure 2. At the end of the survey, users were optionally asked to give any general feedback in free text form.

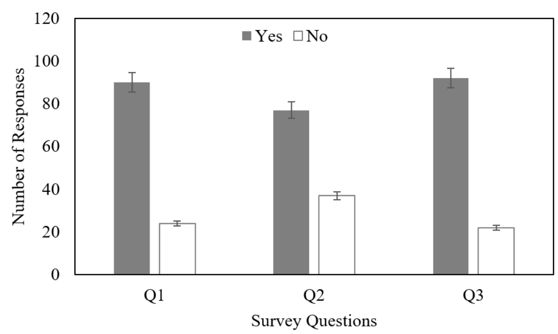


Fig. 2: Summary of Veracity Score Survey Responses

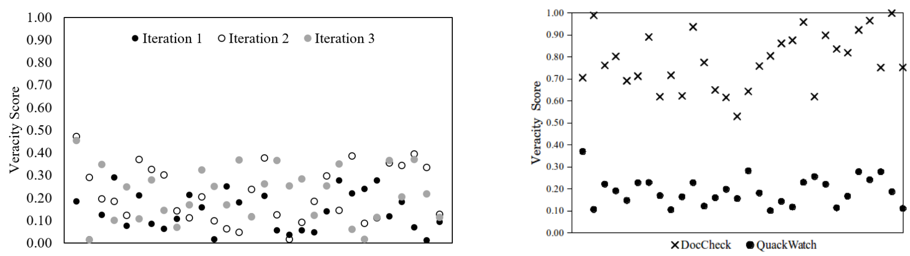
4.6 Veracity Score on Unproven Cancer Treatments

We randomly selected 30 articles on cancer from QuackWatch[12], a web site indexing unproven treatments [11]. These selected articles were input to MedFact to compute a veracity score in order to determine whether the score would align with experts' opinions. The veracity score for the selected articles is summarized in Figure 3(a), showing an overall low score for the articles, hence a consensus with the opinions of the experts who identified the unproven claims, as well as comparable results with the study on predicting unproven cancer treatments by [11].

4.7 Online Medic Discussions Evaluated via Veracity Score

To further evaluate the performance and representative accuracy of MedFact's veracity score, we randomly sampled 30 answers posted on the DocCheck forums[13]. DocCheck allows verified medical professionals to ask questions and post answers. The results showed an average veracity score of 78.32%. A comparison of the veracity scores for medic posts versus QuackWatch averaged scores is presented in Figure 3(b).

[12] QuackWatch web site `http://quackwatch.org`
[13] DocCheck web site `http://doccheck.com`



(a) QuackWatch (b) DocCheck v. QuackWatch

Fig. 3: Veracity Score Comparisons

5 Discussion

Regarding the results of the survey, users did provide generally positive feedback to all three questions. However, regarding the accuracy of the veracity score, users gave less than expected positive feedback. Further analysis into the responses revealed that this was related to the second posting on apricot pits as a cure for cancer, accounting for 70.27% of the lower positive feedback. We investigated the free text feedback to further understand user perspectives. We discovered that the majority of survey participants viewed apricot pit treatments as a homeopathic remedy that should not be covered by scientific literature. Overall, the survey recorded positive feedback from 67.54% of the responses regarding the veracity score accuracy. One area of improvement was the phrase pairs used for our CNN training data, which were limited in number based on the HSE data set. Currently, to the best of our knowledge, there are no medical data sets containing phrase pairs that are annotated for agreement, and involvement from medical experts is essential in order to improve the quality of the phrase pairs. Regarding the comparison between DocCheck and QuackWatch, the results were as expected, with the DocCheck veracity scores being significantly higher than QuackWatch. Moreover, a binary clustering effect can be observed between credible and untrustworthy posts by use of veracity score.

6 Conclusion

The modern patient desires self-education of medical concepts, and seeks to be part of the diagnosis process. However, there is a communication barrier when dealing with technical medical terminology, which could have led to patients seeking more personal and narrative-based sources of medical information. MedFact is an initial step towards bringing trusted and patient-friendly medical knowledge into social media discourse. In this study, we addressed the challenge of veracity in online health information by automating the evidence-based medicine methodology, thereby incorporating medical knowledge into social media discussions while taking into account layperson terminology and hierarchy of evidence.

We also demonstrated the MedFact algorithm, which models trustworthiness and reliability of online information using machine learning and facilitates recommendation of trusted medical articles, ultimately empowering online users to make informed decisions about health information they are consuming. Preliminary work towards a granular veracity score was demonstrated using practical and systematic state-of-the-art methods from deep learning, information retrieval, and text processing. We performed an in-depth experimental analysis of the accuracy and performance of MedFact using the Health Stack Exchange, QuackWatch, and DocCheck data sets via metrics such as precision and recall, as well as qualitative analysis via survey. MedFact improves upon existing approaches towards determining credibility in health-related social media, where ratings, user reputations, and "wisdom of the crowd" are being used predominantly. Firstly, ratings and reputations are more subjective than cross-referencing trusted medical literature. Secondly, ratings require extensive community input, but MedFact relies on existing and comprehensive medical knowledge bases available via the TRIP database. Thirdly, MedFact presents a granular approach that computes veracity from the phrase-level to the document-level. For future work, an interesting aspect to explore is that of agreement between trusted medical articles. Various medical articles may occasionally contain contradictory facts which need to be resolved before incorporating these facts into the MedFact algorithm. A self-referencing veracity score between two trusted phrases can be determined using truth discovery algorithms.

Acknowledgement

We thank the Alberta Machine Intelligence Institute (Amii) for funding this research.

References

1. Kata, A.: 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. Rippen, H., Risk, A.: e-Health Code of Ethics (May 24). Journal of Medical Internet Research **2**(2) (May 2000)
3. Greenhalgh, T.: How to Read a Paper: The Basics of Evidence-Based Medicine. John Wiley & Sons (2010)
4. Ackley, B.J.: Evidence-Based Nursing Care Guidelines: Medical-Surgical Interventions. Elsevier Health Sciences (2008)
5. Child, J.: Trust-The Fundamental Bond in Global Collaboration. Organizational Dynamics **29**(4) (2001) 274–288
6. Varlamis, I., Eirinaki, M., Louta, M.: 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. (2010) 168–175
7. Abdaoui, A., Azé, J., Bringay, S., Poncelet, P.: Collaborative Content-Based Method for Estimating User Reputation in Online Forums. In: Web Information Systems Engineering. Springer (2015) 292–299
8. Grant, S., Betts, B.: Encouraging User Behaviour with Achievements: An Empirical Study. IEEE International Working Conference on Mining Software Repositories (MSR) (2013) 65–68

9. Aljazzaf, Z.M.: Trust-Based Service Selection. PhD thesis, University of Western Ontario (2011)
10. Park, M.: HealthTrust: Assessing the Trustworthiness of Healthcare Information on the Internet. PhD thesis, University of Kansas (2013)
11. Aphinyanaphongs, Y., Aliferis, C., et al.: Text Categorization Models for Identifying Unproven Cancer Treatments on the Web. In: World Congress on Medical Informatics (MedInfo), IOS Press (2007) 968
12. Oliphant, T.: "I Am Making My Decision On The Basis Of My Experience": Constructing Authoritative Knowledge about Treatments for Depression. Canadian Journal of Information and Library Science **33**(3-4) (2009) 215–232
13. Stephens, G.J., Silbert, L.J., Hasson, U.: Speaker–Listener Neural Coupling Underlies Successful Communication. Proceedings of the National Academy of Sciences **107**(32) (2010) 14425–14430
14. Nyhan, B., Reifler, J., Richey, S., Freed, G.L.: Effective Messages in Vaccine Promotion: A Randomized Trial. Pediatrics (2014)
15. Nyhan, B., Reifler, J.: When Corrections Fail: The Persistence of Political Misperceptions. Political Behavior **32**(2) (2010) 303–330
16. Plous, S.: The Psychology of Judgment and Decision Making. McGraw-Hill (1993)
17. Dunning, D.: The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. Advances in Experimental Social Psychology **44** (2011) 247
18. Proctor, R., Schiebinger, L.L.: Agnotology: The Making and Unmaking of Ignorance. Stanford University Press (2008)
19. Henderson, J.: Expert and Lay Knowledge: A Sociological Perspective. Nutrition & Dietetics **67**(1) (2010) 4–5
20. Straus, S.E., Richardson, S.W., Glasziou, P., Haynes, B.R.: Evidence-Based Medicine: How to Practice and Teach EBM. Elsevier/Churchill Livingstone (2005)
21. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Text. In: EMNLP. Volume 4. (2004) 404–411
22. Cornet, R., de Keizer, N.: Forty Years of SNOMED: A Literature Review. BMC Medical Informatics and Decision Making **8**(1) (2008) S2
23. Smith, C., Stavri, P.: Consumer Health Vocabulary. Consumer Health Informatics (2005) 122–128
24. Corcoglioniti, F., Rospocher, M., Aprosio, A.P.: Extracting Knowledge from Text with PIKES. In: International Semantic Web Conference. (2015)
25. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv (2013)
26. Brassey, J.: TRIP Database: Identifying High Quality Medical Literature from a Range of Sources. New Review of Information Networking **11**(2) (2005) 229–234
27. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
28. Pang, B., Lee, L., et al.: Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval **2**(1–2) (2008) 1–135
29. De Marneffe, M.C., Manning, C.D.: Stanford Typed Dependencies Manual. Technical report, Technical Report, Stanford University (2008)
30. Johnson, R., Zhang, T.: 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). (2015)
31. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics (1977) 159–174