**BMC Medical Informatics and Decision Making**

**RESEARCH ARTICLE**                                                                                 **Open Access**

# AutoDiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks

Laura Kinkead[1,2], Ahmed Allam[1,2*] 📖 and Michael Krauthammer[1,2,3]

## Abstract

**Background:** Patients increasingly turn to search engines and online content before, or in place of, talking with a health professional. Low quality health information, which is common on the internet, presents risks to the patient in the form of misinformation and a possibly poorer relationship with their physician. To address this, the DISCERN criteria (developed at University of Oxford) are used to evaluate the quality of online health information. However, patients are unlikely to take the time to apply these criteria to the health websites they visit.

**Methods:** We built an automated implementation of the DISCERN instrument (Brief version) using machine learning models. We compared the performance of a traditional model (Random Forest) with that of a hierarchical encoder attention-based neural network (HEA) model using two language embeddings, BERT and BioBERT.

**Results:** The HEA BERT and BioBERT models achieved average F1-macro scores across all criteria of 0.75 and 0.74, respectively, outperforming the Random Forest model (average F1-macro = 0.69). Overall, the neural network based models achieved 81% and 86% average accuracy at 100% and 80% coverage, respectively, compared to 94% manual rating accuracy. The attention mechanism implemented in the HEA architectures not only provided 'model explainability' by identifying reasonable supporting sentences for the documents fulfilling the Brief DISCERN criteria, but also boosted F1 performance by 0.05 compared to the same architecture without an attention mechanism.

**Conclusions:** Our research suggests that it is feasible to automate online health information quality assessment, which is an important step towards empowering patients to become informed partners in the healthcare process.

**Keywords:** Machine learning, Information quality, Natural language processing, Neural networks, Health communication

## Background

Patients often turn to search engines and online content before, or in place of, talking with a health professional [1]. However, online health information is not regulated, and prior studies have found wide variations in information quality [2]. Poor risk communication, biased writing, and lack of transparency about the source of the information plague online health texts [3, 4]. This presents a real risk to patients, in the form of misinformation [5–7] and negatively affecting their interactions with health care providers [8, 9].

In response to this problem, many organizations, such as the Health on the Net Organization, the Journal of the American Medical Association, and the National Health Service of the UK, have established guidelines for assessing the quality of online health information [10]. These guidelines describe a set of criteria an article must meet to be considered of high quality. It is worth noting that *quality* is distinct from *accuracy*. While these guidelines check

*Correspondence: ahmed.allam@uzh.ch
[1]Department of Quantitative Biomedicine, UZH, Schmelzbergstrasse 26, Zurich, Switzerland
[2]Biomedical Informatics, University Hospital of Zurich, Zurich, Switzerland
Full list of author information is available at the end of the article

for indicators of well written, unbiased, and evidence based articles, they do not attempt to verify the scientific accuracy of the information (a significantly more challenging problem). Similarly, the concept of quality is also distinct from that of *credibility*, or how likely a reader is to believe the information. The propensity to which readers believe the content they consume is influenced not only by information accuracy, but also structural aspects of the media, such as a website's design, appearance, and overall readability [11]. Thus, quality guidelines form a basis by which systems may affect individual's perceptions of credibility, without breaching into the field of information accuracy assessment.

The implementation strategies of these quality guidelines so far fall into two categories: Distributed Guidelines and Centralized Approvers. However, both of these strategies have scalability issues that limit their reach and prevent them from broadly affecting patient information consumption [10]. In the following section, we describe both of these implementation approaches in use today, and then describe a third solution that addresses the issue of scalabilty.

**Distributed Guidelines** One approach to helping patients find high quality health information is to develop a criteria and publish it as a public tool citizens can use. An example of this approach is the DISCERN instrument [12]. The DISCERN instrument's criteria are specifically designed to be able to be understood and applied by any lay person; no medical knowledge is required. This implementation approach puts a significant burden on the patient. For this approach to be successful, the patient has to be aware of the guideline, learn how to evaluate the criteria, and take considerable time to apply the guidelines to every website the patient encounters.

**Centralized Approvers** The second implementation approach in use today is Centralized Approvers. In this approach, an organization manually assesses web pages for health information quality. An example of this approach is the Health on the Net Foundation, which developed the HONcode guidelines. It assesses websites for quality, and allows those that pass their criteria to display a HONcode badge on their webpage [13]. A variant on this approach is to register all manually approved content in a centralized repository. Patients can search the repository with the confidence that all listed sites have been vetted for quality.

The Centralized Approver approach is not scalable in the face of a massive and rapidly growing internet. Quality assessment is a costly manual process. Not only do new pages need to be evaluated, but previously-evaluated pages need to be re-evaluated on a regular basis in case of content changes [10].

**Automated Assessment** An automated quality assessment process is key to providing the public with scalable tools for assessing online health information quality.

Initial attempts to automate the assessment of health information used simplistic approaches, such as readability scores, and did not capture more complex issues with health information, such as tone and bias [4]. A machine learning model developed by the HON organization showed promising but limited initial results [14]. But with the recent developments in machine learning and natural language processing methods, there is a renewed opportunity for tackling this problem. Neural Language Models have been successfully applied in many domains, including translation, question answering, and many more [15–18], capturing details and nuances in language that made information quality assessment an expensive manual process for so long.

### Research objectives

In this research, we study and develop machine learning models to automate the application of the DISCERN instrument. The DISCERN instrument was developed by Charnock et al. [12] at Oxford University and funded by the National Health Service (UK). The instrument consists of 15 questions to help a lay-person to evaluate the quality of online health information regarding treatment options. The validity of the DISCERN instrument has been evaluated in multiple studies, and is commonly used among researchers [19]. The DISCERN instrument suffers from the same sustainability issues as all distributed guidelines do: patients are unlikely to take the time to apply this criteria to each website they find. In this study, we built and evaluated machine learning models for the automated annotation of the Brief DISCERN criteria [20]. We focus on the Brief DISCERN criteria [20], which is a 6 question subset of the DISCERN crieria that has been shown to capture the quality of health information as reliably as the complete DISCERN instrument. Separate models were developed and tested for each of the five Brief DISCERN questions (one question, Q13, was excluded due to low interrater reliability). We compared the use of traditional machine learning (Random Forest) with feature engineering vs. hierarchical encoder attention-based neural network (HEA) models. We also compared the performance of neural models with the attention mechanism (HEA) and without it (HE). Additionally, for both neural architectures, we experimented with the use of two pre-trained neural language models BERT [21] and BioBERT [22] as embeddings in the HEA and HE models. Thus, in total, we trained and compared 5 different architectures: RF, HE+BERT, HE+BioBERT, HEA+BERT, and HEA+BioBERT.

Kinkead *et al. BMC Medical Informatics and Decision Making* (2020) 20:104

Page 3 of 13

**Table 1** Description of the dataset by health topic

| Topic | Breast Cancer | Arthritis | Depression |
|---|---|---|---|
| Number of Articles | 79 | 88 | 102 |
| Number of Sentences | 10,170 | 10,950 | 13,790 |
| Number of Tokens | 125,891 | 129,759 | 160,597 |
| Avg Sentences per Article | 129 | 124 | 135 |
| Avg Tokens per Article | 1,549 | 1,475 | 1,574 |
| | | | |
| Positive Class Prevalence | | | |
| Q4: References | 13% | 14% | 14% |
| Q5: Date | 20% | 26% | 24% |
| Q9: How Treatment Works | 85% | 28% | 52% |
| Q10: Treatment Benefits | 89% | 80% | 65% |
| Q11: Treatment Risks | 63% | 16% | 33% |

## Methods

### Data collection

Using Google Trends, we identified breast cancer, arthritis, and depression as medical topics with the highest search volume since 2004. Using Google and Yahoo search engines, we identified a total of 269 Web pages (HTML articles) with a focus on treatment choices and options across the 3 topics. Two raters (master's students) were trained for 2 months on using the DISCERN instrument and scoring platform. Both raters scored all articles on DISCERN's 5 point scale. Interrater agreement for the DISCERN criteria was adequate to high, ranging between 0.61–0.91 as measured by the Krippendorf score. The process of building the training corpus is described in more detail in [23]. The dataset is described in Table 1.

### Data preprocessing

We converted the scores for each question in the DISCERN instrument, which ranges from 1-5, into a binary classification, where score 3-5 is passing and score 1-2 is failing the criteria. The texts from the HTML articles were extracted and cleaned using the beautifulsoup library[1].

### Neural network model

We designed and implemented a Hierarchical Encoder Attention-based (HEA) model in PyTorch [24] taking into consideration the structure of our problem and the limits of our training data. The model architecture design is primarily motivated by the intrinsic hierarchy of the documents (i.e. sequences of word/tokens represent a sentence, and sequences of sentences represent a document). In addition, our attention-based modeling architecture reflects the property that passing or failing the DISCERN criteria depends on only small fragments throughout the

---

[1]https://www.crummy.com/software/BeautifulSoup/.

article. This architecture enables the model to "pay attention" to single sentences within a larger article.

HEA's architecture is composed of a hierarchical structure with two encoders and a classifier (Fig. 1). The first encoder is a sentence encoder *SentEncoder* (Fig. 2) which is based on a bidirectional recurrent neural network (RNN) that encodes each sentence (i.e. sequence of tokens) into a dense vector representation. The second encoder is a document encoder *DocEncoder* (Fig. 3) which is also based on a bidirectional RNN that encodes the sequence of sentences' representation (i.e. vectors computed from the first encoder) and uses an attention mechanism [15] along with a global context vector to compute a dense vector representation for the whole document. A decoder/classifier maps the document's learned vector representation to the labels using an affine map followed by *softmax* layer computing a probability distribution on the labels for the processed document. An overview of the HEA model architecture can be found in Fig. 1.

### Sentence encoder (SentEncoder)

Formally, given an input sentence $\underline{S} = [\overline{w}_1, \overline{w}_2, \cdots, \overline{w}_{T_S}]$ where $\overline{w}_t$ represents the token representation at position $t$ (i.e. 1-of-$K$ encoding where $K$ is the size of vocabulary $V$ – the set of all tokens in the training corpus), a vanilla RNN will compute a hidden vector at each position (i.e. state vector $\overline{h}_t$ at position $t$), representing a history or context summary of the sequence using the input and hidden states vector form the previous steps. Equation 1 shows the computation of the hidden vector $\overline{h}_t$ using the input $\overline{w}_t$ and the previous hidden vector $\overline{h}_{t-1}$ where $\phi$ is a nonlinear transformation such as $ReLU(z) = max(0, z)$ or $tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$.

$$\overline{h}_t = \phi(\mathbf{W}_{hw}\overline{w}_t + \mathbf{W}_{hh}\overline{h}_{t-1} + \overline{b}_{hw}) \qquad (1)$$

$\mathbf{W}_{hh} \in \mathbb{R}^{D_h \times D_h}$, $\mathbf{W}_{hw} \in \mathbb{R}^{D_h \times D_w}$, $\overline{b}_{hw} \in \mathbb{R}^{D_h}$, represent the RNN's weights to be optimized and $D_h$, $D_w$ are the dimensions of $\overline{h}_t$ and $\overline{w}_t$ vectors respectively. Note that the weights are shared across the network and $D_w$ could be equal to $K$ the size of the vocabulary (i.e. in case of 1-of-$K$ encoding) or the size of a dense embedding vector generated using a language model such as BERT [21]. The use of RNN allows the model to learn long-range dependencies where the network is unfolded as many times as the length of the sequence (sentence in our case) it is modeling. Although RNNs are capable of handling and representing variable-length sequences, in practice, the learning process faces challenges due to the vanishing/exploding gradient problem [25–27]. In this work, we used gated recurrent unit (GRU) [28, 29] to overcome the latter challenges by updating the computation mechanism of the

Kinkead *et al. BMC Medical Informatics and Decision Making* (2020) 20:104

Page 4 of 13
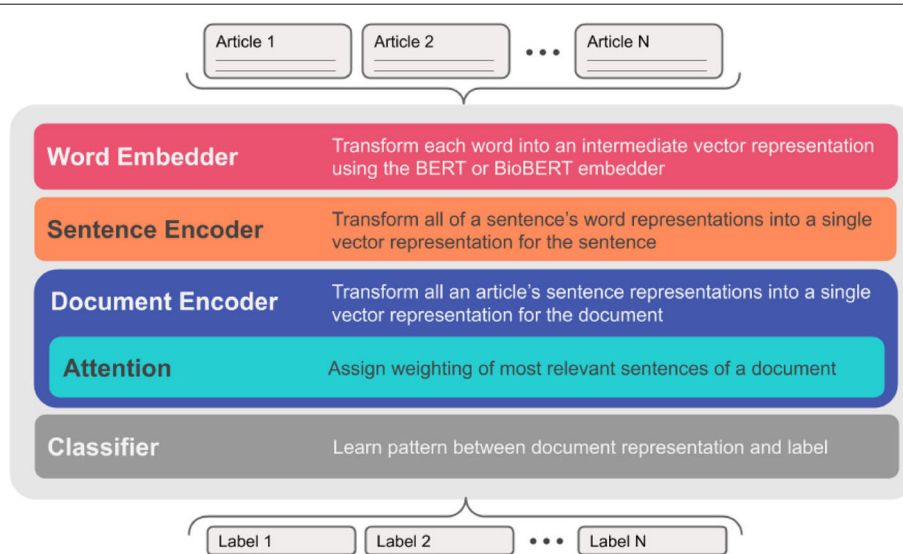


**Fig. 1** Overview of the HEA neural network architecture. In lieu of traditional feature engineering, the HEA architecture learns representations at the word, sentence, and document level before making a classification. Word representations are generated by the pre-trained BERT embedder. An attention mechanism aids in learning a document representation from amongst many sentences

hidden state vector $\overline{h}_t$ through the specified equations below.

$$\overline{z}_t = \sigma(\mathbf{W}_{hw}^z \overline{w}_t + \mathbf{W}_{hh}^z \overline{h}_{t-1} + \overline{b}_{hw}^z) \quad \text{(update gate)}$$

$$\overline{r}_t = \sigma(\mathbf{W}_{hw}^r \overline{w}_t + \mathbf{W}_{hh}^r \overline{h}_{t-1} + \overline{b}_{hw}^r) \quad \text{(reset gate)}$$

$$\overline{\tilde{h}}_t = \phi(\mathbf{W}_{hw}^{\tilde{h}} \overline{w}_t + \overline{r}_t \odot \mathbf{W}_{hh}^{\tilde{h}} \overline{h}_{t-1} + \overline{b}_{hw}^{\tilde{h}}) \quad \text{(new state/memory cell)}$$

$$\overline{h}_t = (1 - \overline{z}_t) \odot \overline{\tilde{h}}_t + z_t \odot \overline{h}_{t-1} \quad \text{(hidden state vector)}$$

The GRU model computes a reset gate $\overline{r}_t$ that is used to modulate the effect of the previous hidden state vector $\overline{h}_{t-1}$ when computing the new memory vector $\overline{\tilde{h}}_t$. The update gate $\overline{z}_t$ determines the importance/contribution of the newly generated memory vector $\overline{\tilde{h}}_t$ compared to the previous hidden state vector $\overline{h}_{t-1}$ when computing the current hidden vector $\overline{h}_t$. The weights $\mathbf{W}_{hw}^z$, $\mathbf{W}_{hw}^r$, $\mathbf{W}_{hw}^{\tilde{h}}$
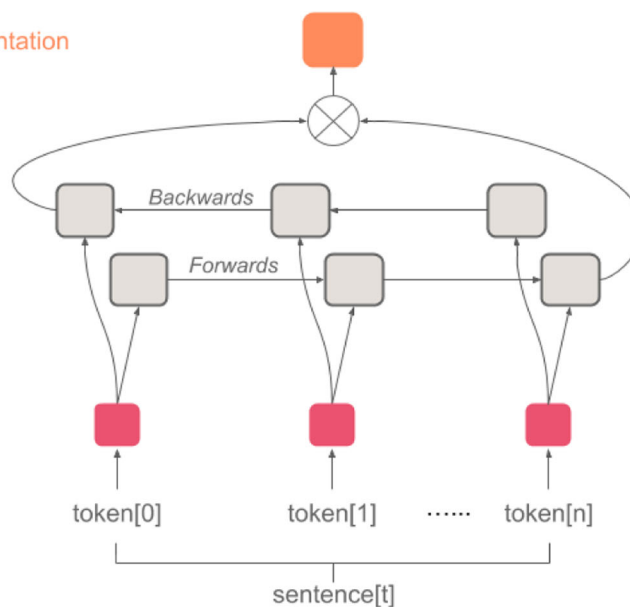


**Fig. 2** HEA's *SentEncoder* architecture for computing sentence embedding
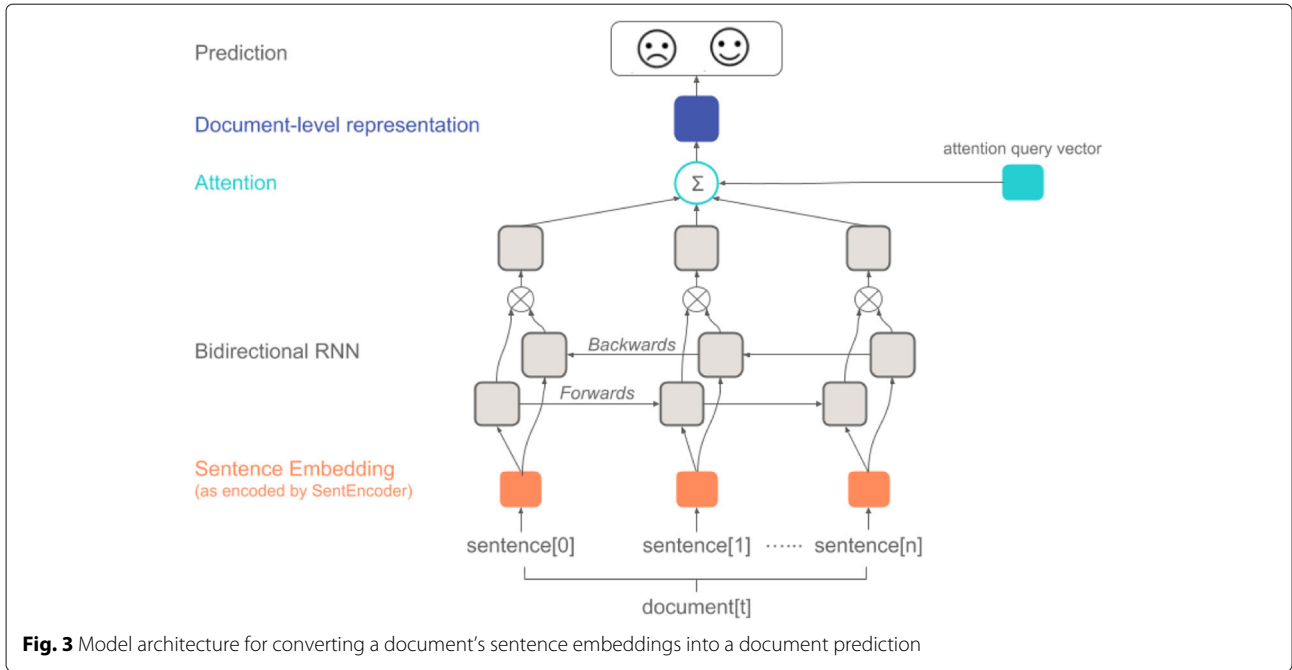
**Fig. 3** Model architecture for converting a document's sentence embeddings into a document prediction

each $\in \mathbb{R}^{D_h \times D_w}$ and $\mathbf{W}_{hh}^z$, $\mathbf{W}_{hh}^r$, $\mathbf{W}_{hh}^{\tilde{h}}$ each $\in \mathbb{R}^{D_h \times D_h}$. The biases $\overline{b}_{hw}^z$, $\overline{b}_{hw}^r$, $\overline{b}_{hw}^{\tilde{h}}$ each $\in \mathbb{R}^{D_h}$ where $D_h$ and $D_w$ are the dimensions of $\overline{h}_t$ and $\overline{w}_t$ vectors respectively. The operator $\sigma$ represents the *sigmoid* function, $\phi$ the *tanh* or *ReLU* function, and $\odot$ the element-wise product (i.e. Hadamard product).

The *SentEncoder* uses a bidirectional GRU that computes two hidden state vectors $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ for each token $\overline{w}_t$ in sentence $\underline{S}$ corresponding to left-to-right and right-to-left GRU encoding of the sentence. We experimented with two options for computing sentence representation vector $\overline{S}$: (1)*concatenation* $[\overrightarrow{h}_{T_S}{}^\top; \overleftarrow{h}_0{}^\top]^\top$, and/or (2) *summation* $[\overrightarrow{h}_{T_S} + \overleftarrow{h}_0]$ of the computed left and right hidden state vectors of the last $\overline{w}_{T_S}$ and first $\overline{w}_0$ tokens respectively in sentence $\underline{S}$.

### Document encoder (DocEncoder) with attention

Originally, each document $\underline{Doc}$ in our corpus is composed of a sequence of sentences (i.e. $\underline{Doc} = [\underline{S}_1, \underline{S}_2, \cdots, \underline{S}_{T_{Doc}}]$ where $\underline{S}_i$ represents the $i^{th}$ sentence and $T_{Doc}$ is the number of sentences in $\underline{Doc}$). Each sentence $\underline{S}_i$ is composed of a sequence of tokens (as described in "Sentence encoder (SentEncoder)" section above) that are processed using *SentEncoder* model to compute the sentence vector representation $\overline{S}_i$. As a result, the processed document $\underline{Doc}^{proc}$ is a sequence of sentences' vector representation (i.e. $\underline{Doc}^{proc} = [\overline{S}_1, \overline{S}_2, \cdots, \overline{S}_{T_{Doc}}]$) that is used as input to *DocEncoder* model. The *DocEncoder* uses a bidirectional GRU that computes two hidden state vectors

$\overrightarrow{l}_i$ and $\overleftarrow{l}_i$ for each sentence representation $\overline{S}_i$ corresponding to left-to-right and right-to-left GRU encoding of the sentences in $\underline{Doc}^{proc}$. We experimented with two options for joining both hidden state vectors $\overrightarrow{l}_i$ and $\overleftarrow{l}_i$ into one vector using: (1)*concatenation* $[\overrightarrow{l}_i{}^\top; \overleftarrow{l}_i{}^\top]^\top$, and/or (2) *summation* $[\overrightarrow{l}_i + \overleftarrow{l}_i]$ that will be denoted by $\overleftrightarrow{l}_i$ from now on. Hence, the output of the *DocEncoder* is a sequence of *joined* hidden state vectors $\underline{O} = [\overleftrightarrow{l}_1, \overleftrightarrow{l}_2, \cdots, \overleftrightarrow{l}_{T_{Doc}}]$ that is fed to an *attention* layer to compute the weights associated with each vector which in turn are used to compute a weighted vector sum to obtain a document vector representation $\overline{z}$.

### Attention layer

For many of the DISCERN criteria, pass or fail of the criteria depends on only small fragments throughout the article. For example, for the question "Is it clear when the information used or reported in the publication was produced?", there is likely only one line among a 200+ sentence article (i.e. "Last reviewed on...") that determines whether the article passes the criteria. Our attention-based modeling architecture reflects this problem structure: the model can "pay attention" to single sentences within a larger article. We adapt the idea of *global* attention model [16] in which a global context/query vector $\overline{q}$ (i.e. trainable parameters in the model) was used along with the output $\underline{O} = [\overleftrightarrow{l}_1, \overleftrightarrow{l}_2, \cdots, \overleftrightarrow{l}_{T_{Doc}}]$ from *DocEncoder*

to generate document representation vector $\bar{z}$. The objective is to compute *attention* weights for every $\overleftrightarrow{l_i}$ vector such that $\bar{z} = \sum_{i=1}^{T_{Doc}} \alpha_i \overleftrightarrow{l_i}$ where $\alpha_i$ is the normalized weight computed using Eq. 2.

$$\alpha_i = \frac{\exp(score(\bar{q}, \overleftrightarrow{l_i}))}{\sum_{k=1}^{T_{Doc}} \exp(score(\bar{q}, \overleftrightarrow{l_k}))} \tag{2}$$

For the attention *scoring* function, we experimented with two options inspired by the *additive* approach [16, 30] and the *scaled dot-product* work in [15] (see Equations 3 and 4 respectively). In Eq. 3, the score is computed using three operations: (1) a weight matrix $\mathbf{W}_{ql}^l \in \mathbb{R}^{D_q \times D_l}$ maps $\overleftrightarrow{l_i}$ to a fixed-length vector of dimension equal to the query vector $\bar{q}$ (i.e. $D_q$), (2) a non-linear transformation *tanh* is applied, and (3) a *dot-product* with $\bar{q}$ is performed. In contrast, in Eq. 4, the score is computed by performing a *dot-product* between the query vector $\bar{q}$ and $\overleftrightarrow{l_i}$ scaled by $D_l$ which is the dimension of both vectors in similar approach to [15]. Our choice of attention score functions from the vast array of options in the literature [16, 30], was based on limiting the number of parameters in our model given the size of our dataset.

$$score(\bar{q}, \overleftrightarrow{l_i}) = \bar{q}^\top tanh(\mathbf{W}_{ql}^l \overleftrightarrow{l_i}) \tag{3}$$

$$score(\bar{q}, \overleftrightarrow{l_i}) = \frac{\bar{q}^\top \overleftrightarrow{l_i}}{\sqrt{D_l}} \tag{4}$$

### Decoder/output classifier

The last layer in the HEA model takes as input the computed document representation vector $\bar{z}$ from the *DocEncoder* layer and performs an affine transformation followed by *softmax* operation to compute a probability distribution on the labels for the document under consideration. That is, the outcome $\hat{y}$ for a given Brief DISCERN criterion is computed using Eq. 5

$$\hat{y} = \sigma(\mathbf{W}_{V_{label}z} \bar{z} + \bar{b}_{V_{label}}) \tag{5}$$

where $\mathbf{W}_{V_{label}z} \in \mathbb{R}^{|V_{label}| \times D_z}$, $\bar{b}_{V_{label}} \in \mathbb{R}^{|V_{label}|}$ represents the classifier's weights to be optimized, $V_{label} \in \{0, 1\}$ is the set of admissible labels for a criterion (binary variable in our case), $|V_{label}|$ is the number of labels, $D_z$ is the dimension of $\bar{z}$ (document representation vector), and $\sigma$ is the *softmax* function. As a result, the outcome $\hat{y}$ represents a probability distribution over the set of possible labels $V_{label}$.

### Objective function

We used cross-entropy loss as our objective function for each Brief DISCERN criterion model. The loss function for a $j^{th}$ document is defined by Eq. 6 where $y_c \in \{0, 1\}$ is equivalent to $\mathbb{1}[y = c]$ (i.e. a boolean indicator equal to 1 when $c$ is the reference/ground-truth class), and $\hat{y}_c$ is the probability of the class $c$. The objective function for the whole training set $D_{train}$ is defined by the average loss across all the documents in $D_{train}$ plus a weight regularization term (i.e. $l_2$-norm regularization) applied to the model parameters represented by $\boldsymbol{\theta}$ (see Eq. 7).

$$l^{(j)} = -\sum_{c=1}^{|V_{label}|} y_c^{(j)} \times log(\hat{y}_c^{(j)}) \tag{6}$$

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{j=1}^{N} l_j + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2 \tag{7}$$

In addition to the $l_2$-norm regularization, we also experimented with dropout [31] by deactivating neurons in the network layers using probability $p_{dropoout}$. Moreover, we used pre-trained language models such as BERT [17, 21] and BioBERT [22] to extract token embeddings that are used as input to HEA's model (i.e. representation of token $\bar{w}_t$).

We additionally implemented a neural-based model (HE) that follows the same architecture of HEA model but without the attention layer such that the output of the *DocEncoder* representing a sequence of *joined* hidden state vectors $\underline{O} = [\overleftrightarrow{l_1}, \overleftrightarrow{l_2}, \cdots, \overleftrightarrow{l_{T_{Doc}}}]$ is mean pooled (i.e. averaged) to obtain an overall document vector representation $\bar{z}$.

### Hyperparameter optimization for neural models

We developed a multiprocessing module that used a uniform random search strategy [32] that randomly chose a set of hyperparameters configurations (i.e. layer depth, embedding size, attention approach, etc.) from the set of all possible configurations. Then the best configuration for each model (i.e. the one achieving best performance on the validation set) was used for the final training and testing.

### Baseline machine learning models

For the traditional modeling approach, the content of each article was converted into a bag of words representation and weighted using the term frequency–inverse document frequency (TF-IDF) weighting scheme. We also computed a set of features based on the existence of HTML links, bibliography keywords, references to medical terms (extracted using MetaMap Lite [33]), and named entities within the text, as well as a measure of text polarity. Recursive feature elimination with cross validation was

used to identify the optimal subset of features. For its ease of interpretability and good performance on feature sets with many categorical variables, we implemented a Random Forest model with scikit-learn [34] to predict if the criterion is fulfilled or not for every criterion in Brief DISCERN.

### Experimental setup

We followed a stratified 5-fold cross-validation scheme where each fold was defined as a distinct 80%-20% train-test split. Due to the imbalance in outcome classes, training examples were weighted inversely proportional to class/outcome frequencies in the training data. Articles from the three health topics were randomly distributed between the 5 folds. Within each fold, parameter selection was performed with a validation set consisting of 10% of the training set. During the training of the models, the epoch in which the model achieved the best F1-macro score on the validation set was recorded, and model state as it was trained up to that epoch was saved. This best model, as determined by the validation set, was then tested on the test split.

Model performance was evaluated using F1-macro and classification accuracy. In this quality assessment problem, we value precision equally with recall, so F1 is a good measure that captures both. The evaluation of the trained models was based on their average performance on the test sets of the five folds.

We also performed a coverage analysis to determine how the model could be adapted to handle uncertainty. In addition to classifying articles as low or high quality, we also have the option of allowing the model report that it is unsure about a criteria. In instances when the model has a low confidence in its prediction, it is more valuable to the user for the model to convey that uncertainty, than to make a less accurate prediction. In addition, there is also the option to send articles where the model has low confidence to a human for manual evaluation. However, there is a direct trade-off between the quality (accuracy) and the quantity the predictions; by requiring a higher threshold of confidence, the model will by definition make a fewer number of predictions. The frequency with which the model makes prediction above a certain confidence threshold, i.e. outputs a prediction to the user, is called coverage. We calculated the models' accuracy at different levels of coverage and their associated confidence thresholds. For example, to calculate the accuracy associated with a coverage of 80%, we computed the 20th percentile prediction confidence score, and computed accuracy metrics on only the articles with prediction confidence scores (i.e. the probability from *softmax* layer) that exceed the 20th percentile. Predictions that are below this threshold are considered *unsure*. These are instances where the model would abstain from making

a prediction, or the article could be sent for manual review.

### Code availability

The data preprocessing and the models' implementation (training and testing) workflow is made publicly available at https://github.com/uzh-dqbm-cmi/auto-discern.

### Results

We compared the performance of the five trained models (Random Forest, HEA with BERT and BioBERT embeddings, and HE with BERT and BioBert embeddings) across all five folds using F1-macro scores (Table 2 and Fig. 4). Overall, the HEA architecture perforemd the best, scoring an average F1-macro score of 0.75 with BERT embeddings and 0.74 with BioBERT embeddings. In comparison, the HE architectures without the attention mechanism averaged 0.70 on both embeddings. The Random Forest model achieved an average F1-macro score of 0.69.

Almost all models performed the best on question 4 ("Is it clear what sources of information were used to compile the publication [other than the author or producer]?") with HEA BERT, HEA BioBERT and Random Forest scoring 0.86, 0.80, and 0.83 respectively. The HE models performed worse on this question, with 0.72 with BERT and 0.71 with BioBERT. All five models achieved high F1-macro scores on question 5 ("Is it clear when the information used or reported in the publication was produced?") with HEA BioBERT coming first (0.82), HE BioBERT second (0.78), HEA BERT and HE BERT tying for third (0.77), and Random Forest last (0.70).

For treatment related questions, HEA BioBert performed the best. On question 9 ("Does it describe how each treatment works?"), HEA BioBERT came first with an average F1-macro score of 0.72, and the remaining models ranging between 0.68 and 0.66). For question 11 ("Does it describe the risks of each treatment?"), both neural models using the BioBERT Embeddings performed the best: the HEA- and HE BioBERT models scored 0.81 and 0.80 respectively, with the remaining models following betwwen 0.76 and 0.72. In contrast, for question 10 ("Does it describe the benefits of each treatment?"), the neural models using the BERT embeddings performed better, with HEA BERT at 0.66 and HE BERT at 0.60, with the remaining models ranging between 0.56 and 0.53. It worth mentioning that Q10 has the greatest class imbalance (i.e. 77% of the articles in the data set described the benefits of the treatment).

As measured by F1-macro, HEA-BioBert took first place in 3 of the 5 questions, and HEA-Bert took first in the remaining 2 questions (see Table 2). However, when computing the average score on all questions, HEA BERT and BioBERT performed comparably, and the variance

**Table 2** Average F1-macro scores with standard deviation by model architecture
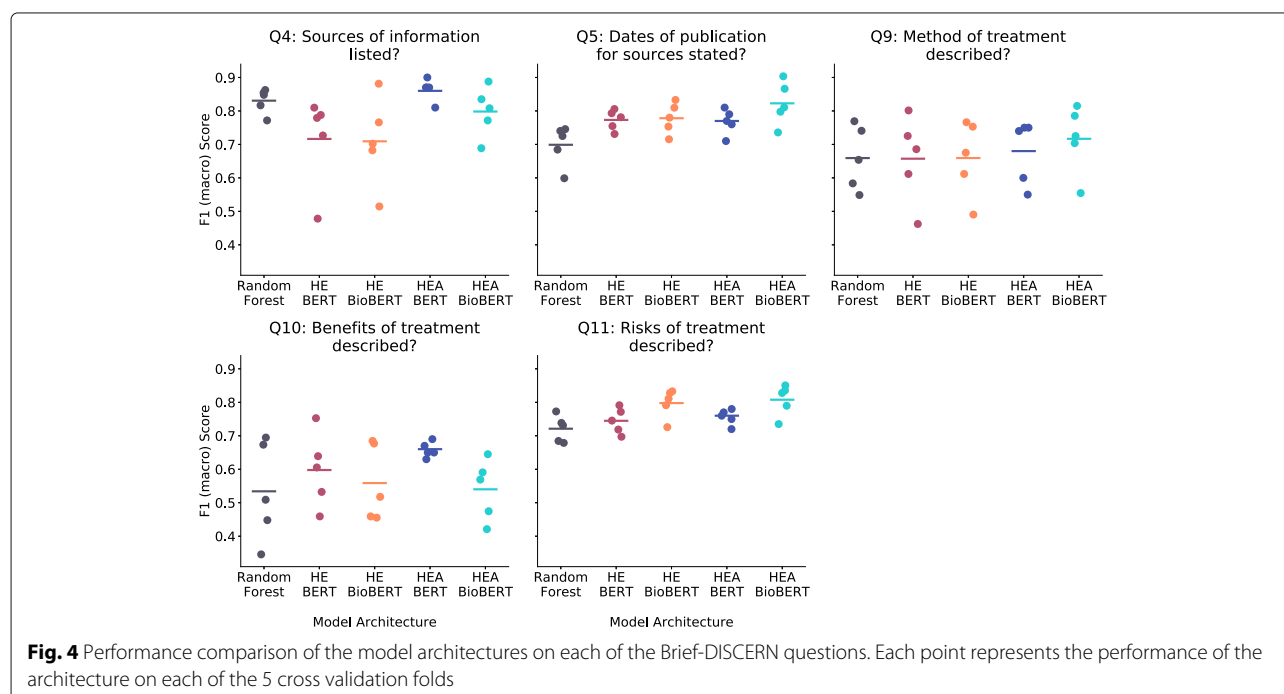
| Model Architecture | Q4: References | Q5: Date | Q9: How Treatment Works |
| --- | --- | --- | --- |
| Random Forest | 0.83 ( 4) | 0.70 ( 6) | 0.66 (10) |
| HE BERT | 0.72 (14) | 0.77 ( 3) | 0.66 (13) |
| HE BioBERT | 0.71 (13) | 0.78 ( 5) | 0.66 (11) |
| HEA BERT | 0.86 ( 3) | 0.77 ( 4) | 0.68 (10) |
| HEA BioBERT | 0.80 ( 7) | 0.82 ( 6) | 0.72 (10) |
| Model Architecture | Q10: Tt. Benefits | Q11: Tt. Risks | All Questions Avg |
| Random Forest | 0.53 (15) | 0.72 ( 4) | 0.69 |
| HE BERT | 0.60 (11) | 0.74 ( 4) | 0.70 |
| HE BioBERT | 0.56 (11) | 0.80 ( 4) | 0.70 |
| HEA BERT | 0.66 ( 2) | 0.76 ( 3) | 0.75 |
| HEA BioBERT | 0.54 ( 9) | 0.81 ( 5) | 0.74 |

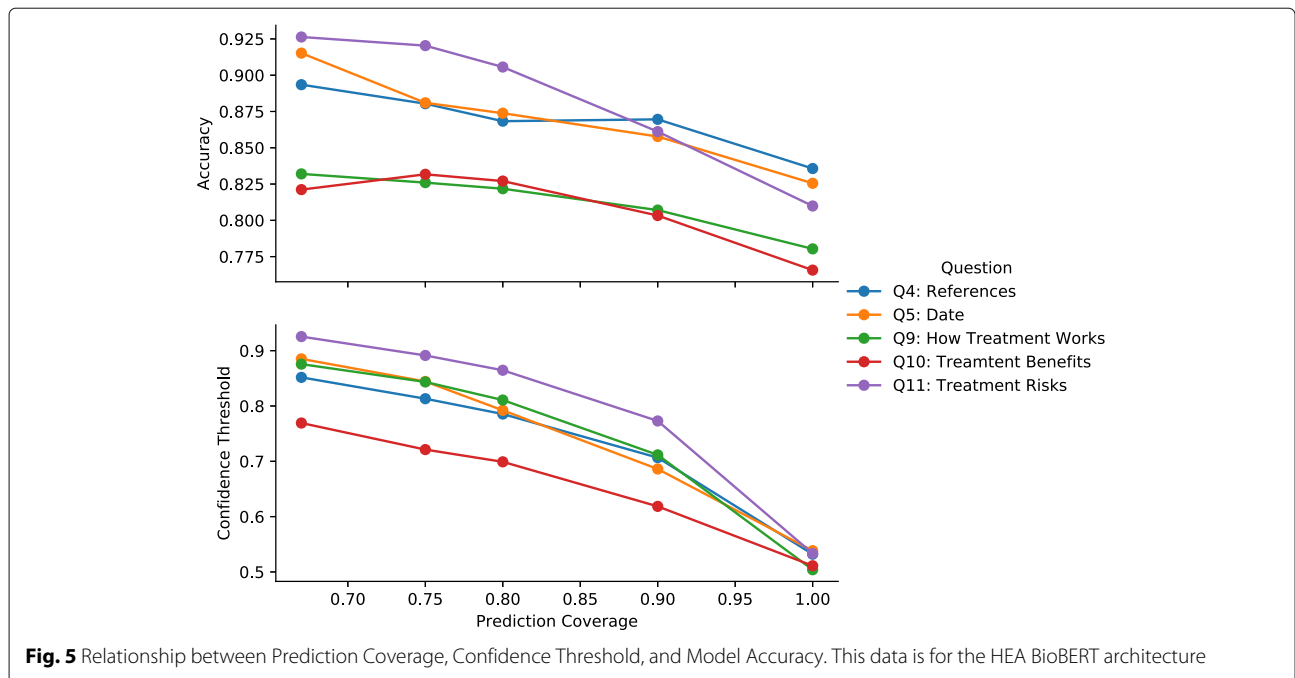across folds was lower for HEA BERT compared to HEA BioBERT (see Fig. 4).

We explored the relationship between model's coverage, accuracy, and confidence (prediction probability threshold) focusing on the BioBERT model (see Fig. 5). The trend is that as the model's coverage decreases, the higher is the confidence (i.e. outcome probability) and the accuracy of the model. At 80% coverage, the model achieves 86% average accuracy with average confidence equal to 0.79 (see Table 3).

Table 4 compares the machine learning model (HEA BioBERT) performance to human manual performance on the DISCERN and HON guidelines. We report the DISCERN Manual Performance as the frequency with which each of our two raters' agree with the aggregated average of both raters (i.e. percent agreement to aggregate). The manual rater accuracy score on our data set averaged 94% (spanning 88% - 97% across criteria). Compared to the DISCERN raters' manual performance of 94%, the models performance was adequate (81% accuracy across all questions at 100% coverage). In order to bring the model's performance closer to the DISCERN raters' manual performance, we could reduce the model coverage to 80%, which would yield an accuracy of 86%.



**Fig. 4** Performance comparison of the model architectures on each of the Brief-DISCERN questions. Each point represents the performance of the architecture on each of the 5 cross validation folds

**Fig. 5** Relationship between Prediction Coverage, Confidence Threshold, and Model Accuracy. This data is for the HEA BioBERT architecture

As an additional comparison, Table 4 also contains HON organization manual percent agreement scores that were computed while developing training sets for their own machine learning models [14]. The average percent agreement among the HON raters was 85% on their full criteria, and this drops to 81% when only considering criteria that are shared with Brief DISCERN (Reference, Date, and Justifiability). Overall, the machine learning model achieved a competitive performance at full coverage compared to HON raters: the model averaged 83% accuracy at 100% coverage on the questions that overlap

with HON, and the HON raters had percent agreement of 81%.

Table 5 shows top-3 sentences (based on attention probability score) belonging to most confidently predicted documents as determined by the prediction probability score for each question in the three medical topics.

Lastly, the time and space requirements for the different architectures were very different. For the neural models, running the full hyperparameter search and training routine took between 25-30 hours on a GPU node with 256 GB of RAM parallelized across 5 Nvidia GTX 1080 GPUs.

**Table 3** Comparison of performance metrics for the HEA BioBERT architecture at 80% and 100% coverage. Coverage refers to the percent of articles the model makes a prediction for (as opposed to abstaining from making a prediction when the model has a confidence below the Threshold). The Precision, Recall, and Accuracy scores reflect the accuracy of the model on the resulting 80% of predicted articles

| Question | Coverage | Threshold | Precision | Recall | Accuracy |
|---|---|---|---|---|---|
| Q4: References | 80% | 0.79 | 0.87 | 0.79 | 87% |
| Q5: Date | 80% | 0.79 | 0.87 | 0.88 | 87% |
| Q9: How Treatment Works | 80% | 0.81 | 0.84 | 0.71 | 82% |
| Q10: Treatment Benefits | 80% | 0.70 | 0.66 | 0.55 | 83% |
| Q11: Treatment Risks | 80% | 0.86 | 0.90 | 0.90 | 91% |
| | | | | | |
| Q4: References | 100% | 0.50 | 0.83 | 0.80 | 84% |
| Q5: Date | 100% | 0.50 | 0.83 | 0.83 | 83% |
| Q9: How Treatment Works | 100% | 0.50 | 0.77 | 0.72 | 78% |
| Q10: Treatment Benefits | 100% | 0.50 | 0.57 | 0.54 | 77% |
| Q11: Treatment Risks | 100% | 0.50 | 0.81 | 0.81 | 81% |

**Table 4** Performance Comparison between 525 Human Manual Rating and Deep Learning Model. Manual performance 526 is reported as percent agreement. Automated performance is reported as Implementation Accuracy (see Table 3).

| Question | Manual Performance | | Automated Performance | |
|---|---|---|---|---|
| | DISCERN | HONcode | HEA BioBERT | |
| | 2 raters | 3 raters | 80% coverage | 100% coverage |
| Q4: References (HoN: Reference) | 96% | 89% | 87% | 84% |
| Q5: Date (HoN: Date) | 88% | 80% | 87% | 83% |
| Q9: How Treatment Works | 92% | | 82% | 78% |
| Q10: Treatment Benefits | 95% | | 83% | 77% |
| Q11: Tt. Risks (HoN: Justifiability) | 97% | 74% | 91% | 81% |
| average | 94% | 81% | 86% | 81% |

In comparison, the baseline model trained in 15 minutes on a machine with 4 CPUs and 16 GB of RAM.

## Discussion

In this research, we developed an attention-based neural network model with the aim to automatically determine the quality of online health information.

The experiments suggest that a neural network model with trained language embeddings on large text corpora (generic or medical) has better performance than a conventional baseline model (Random Forest). Importantly, this superior performance was achieved without the need to hand-craft input features, as was the case with the baseline model. However, it is worth noting that this comes at the trade-off of much higher computing requirements for the neural network models.

Our results reiterate the success of using trained language models [15, 17], and transfer learning [18] in achieving competitive results even on small datasets (as in our case). The BioBERT embeddings show a slight advantage in comparison to BERT ones (Table 2), and we believe this could be due to the medical topics and the language used to describe treatments in each topic.

Our results suggest that the neural attention mechanism not only provided a performance boost over a mean pooled neural architecture, but also enabled greater model explainability. The HEA models performed 7% higher in F1-macro compared to the HE models (Table 2). In addition, inspecting Table 5, it can be seen that the HEA BioBERT model provided reasonable context sentences (i.e. sentences supporting a prediction). In other words, the models identify textual snippets (surrogates, or proxy) in the articles. For question 4 (References), the model identified sentences containing citations, and in question 5 (Date), the model identified text referring to dates when the article was "reviewed", "revised", or "updated". Questions 9 (How Treatment Works) and 10 (Treatment Benefits) are related questions, and we see that the textual snippets identified by the model for these questions overlap, as

expected. However, question 10 achieved poorer accuracy scores, which is probably due to class imbalance in the training data for that question (only 33% of articles were in the negative class). For question 11 (Treatment Risks), the model often identified section headings containing the phrase "side effects".

We compared the models' quality assessment performance to humans manually performing the same task. Our raters achieved an average of 94% agreement across criteria. Similarly, the HON organization reported an average percent agreement of 85% on their criteria, and this drops to 81% when only considering criteria that are shared with Brief DISCERN. While the HEA BioBERT model performed lower than manual raters used in this study (81% vs. 94%), it showed competitive results when compared to HON raters (83% vs. 81% average accuracy). Restricting the HEA BioBERT model's prediction coverage to 80%, we could further improve the prediction performance achieving 86% average accuracy across all criteria. In this case, the model refrains from making a prediction when its prediction probability score, or confidence, is below the 20th percentile. This model or a similarly trained model could be effectively used for pre-screening health web pages and for assisting manual raters in the quality assessment task. As suggested by the HON organization, assisting manual rating with automated systems could reduce manual effort [14].

## Future work

We are seeking to further improve our models' performance to more closely achieve human performance in assessing online health information quality. One straightforward approach is to train on a larger data set using the same model architecture. To achieve this aim, we could look beyond manually labeling more health articles. For example, we could construct a larger corpora by combining our current dataset with other existing bodies of online health information that have independently been assessed as being of high quality. For example, articles

**Table 5** Example sentences that the models paid the most attention to for each disease category. These are the sentences with the highest attention weight for the top three most confidently predicted documents as determined by the prediction probability score. These results are from the HEA BioBERT model.

**Question 4: Sources**

| Breast Cancer | Arthritis | Depression |
| --- | --- | --- |
| 2010 Aug 10;28(23):3784-96. | Nat Rev Rheumatol. | J Abnorm Psychol. |
| 2008;148(5): 358-69. | Leuk Res. | J Abnorm Psychol. |
| Lancet 2007; 369(9555):29–36. | Kelley's Textbook of Rheumatology. | American Journal of Geriatric Psychiatry. |

**Question 5: Date**

| Breast Cancer | Arthritis | Depression |
| --- | --- | --- |
| Review Date: 11/17/2012. | Review Date: 9/26/2011. | Review Date: 3/8/2013. |
| Last Revised: 10/01/2013. | All rights reserved. | All rights reserved. |
| Review Date: 6/5/2012. | Article updated: 31 October 2012. | Page last updated: 1-Oct-2013. |

**Question 9: How Treatment Works**

| Breast Cancer | Arthritis | Depression |
| --- | --- | --- |
| During this surgery, the surgeon removes the axillary lymph nodes as well as the chest wall muscle in addition to the breast. | In this surgery, the healthcare provider actually removes the inflamed synovial tissue. | The basis of this therapy is that behaviours such as inactivity and ruminating on certain thoughts can be key factors in maintaining depression. |
| Radiation therapy is typically done using a large machine that aims the energy beams at your body (external beam radiation). | One part of such therapy involves working with a physical therapist to perform dedicated exercises for muscle strengthening, increasing range | Gentler martial arts which focus on internal control, breathing and mental discipline can be especially useful for combating depressed thinking and improving relaxation skills. |
| Three-dimensional conformal radiation therapy (3D-CRT): As part of this treatment, special computers create detailed three-dimensional pictures | Hydrotherapy differs from swimming because it involves special exercises that you do in a warm-water pool. | Psychoanalytic therapists rely on suggestion, hypnosis, and reeducation to reform self-esteem, and helps the person construct coping strategies to deal with grief, sadness, disappointment, achievement, and pleasure. |

**Question 10: Treatment Benefits**

| Breast Cancer | Arthritis | Depression |
| --- | --- | --- |
| Treating early breast cancer. | Getting Established on DMARD Therapy. | Cognitive Behavioral Therapy for Depression. |
| Targeted therapy for breast cancer. | Medications will not JIA; rather they can help to symptoms and keep disease activity under . | The mindfulness approach uses meditation, yoga, and breathing exercises to focus awareness on the present moment and break negative thinking |
| Adjuvant and Neoadjuvant Therapy for Breast Cancer. | Treatment for Juvenile Rheumatoid Arthritis. | CBT is based on two specific tasks: cognitive restructuring, in which the therapist and patient work together to change thinking patterns, 192 and behavioral activation – in which patients learn to overcome obstacles to participating in enjoyable activities. |

**Question 11: Treatment Risks**

| Breast Cancer | Arthritis | Depression |
| --- | --- | --- |
| The side effects vary depending on which biological therapy drug you have. | Risks: Always talk to your doctor or pharmacist before taking NSAIDs as they may cause serious side effects compared to paracetamol. | Side Effects of ECT . |
| Side effects . | Risks: Always talk to your doctor or pharmacist before taking NSAIDs as they may cause serious side effects compared to paracetamol. | Common side effects of SSRIs include:. |
| Are there side effects or risks from hormone therapy? | Common side effects include a rise in blood pressure, increased hair growth, increased swelling of the gums and an increased risk of developing an infection. | What Are the Risks? |

approved by HON could be used as positive examples in an augmented training set. An additional avenue is to use semi-supervised learning and unsupervised data augmentation approaches [35, 36] where unlabeled data is incorporated to improve classification performance without additional annotation burden.

In future experiments, we plan to further develop our use of language embeddings. For example, in this research we simply used the last layer embeddings from the BERT and BioBERT models. However, recent experiments suggest using different layers (the BERT network contains 12 layers) or further training the embedding networks can yield performance improvements.

Finally, the DISCERN instrument is designed to be applied to articles describing treatment options. Thus, our model's applicability is limited to these types of articles. Similarly, our model does not extend to the medium of social media, which online users are increasingly using to share and consume health information [11]. More research is needed to develop models for assessing the quality of other types and mediums of health information.

## Conclusion

Our study demonstrates that neural models are able to perform online health information quality assessment in accordance with an existing quality criteria (Brief DIS-CERN) with a performance above 80% accuracy. The neural approach achieves a better performance than a conventional approach using Random Forest. In addition, we observe that existing biomedical language models improve performance on this task. Finally, we show that attention-based neural approaches are able to retrieve relevant supporting sentences from the text, which makes model decisions more explainable to users.

### Abbreviations
HE: Hierarchical Encoder Network; HEA: Hierarchical Encoder Network with Attention; HON: Health on the Net Organization; RF: Random Forest; TF-IDF: term frequency–inverse document frequency

### Author's contributions
LK and AA worked on the development of processing and analysis workflow, algorithms and models implementation. LK, AA and MK analyzed and interpreted the data. LK and AA drafted the manuscript. AA and MK supervised and edited the manuscript. All authors approved the final article.

### Availability of data and materials
The preprocessing and the models' implementation (training and testing) workflow is made publicly available at https://github.com/uzh-dqbm-cmi/auto-discern

### Ethics approval and consent to participate
Not Applicable

### Consent for publication
Not Applicable …

### Competing interests
The authors declare that they have no competing interests.

### Author details
[1]Department of Quantitative Biomedicine, UZH, Schmelzbergstrasse 26, Zurich, Switzerland. [2]Biomedical Informatics, University Hospital of Zurich, Zurich, Switzerland. [3]Department of Pathology, Yale University School of Medicine, New Haven, USA.

### References
1. Hesse BW, Nelson DE, Kreps GL, Croyle RT, Arora NK, Rimer BK, Viswanath K. Trust and Sources of Health Information. Arch Intern Med. 2005;165(22):2618. https://doi.org/10.1001/archinte.165.22.2618.
2. Fahy E. Quality of patient health information on the internet: reviewing a complex and. Australas Med J. 2014;7(1):24–8. https://doi.org/10.4066/AMJ.2014.1900.
3. Zhang Y, Sun Y, Xie B. Quality of health information for consumers on the web: A systematic review of indicators, criteria, tools, and evaluation results. J Assoc Inf Sci Technol. 2015;66(10):. https://doi.org/10.1002/asi.23311.
4. Saunders CH, Petersen CL, Durang M-A, Bagley PJ, Elywn G. Bring on the Machines: Could Machine Learning Improve the Quality of Patient Education Materials? A Systematic Search and Rapid Review. JCO Clin Cancer Inform. 2018;2:1–16. https://doi.org/10.1200/cci.18.00010.
5. Murray E, Lo B, Pollack L, Donelan K, Catania J, Lee K, Zapert K, Turner R. The Impact of Health Information on the Internet on Health Care and the Physician-Patient Relationship: National U.S. Survey among 1.050 U.S. Physicians. J Med Internet Res. 2003;5(3):17. https://doi.org/10.2196/jmir.5.3.e17.
6. Allam A, Schulz PJ, Nakamoto K. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating Google output,. J Med Internet Res. 2014;16(4):100. https://doi.org/10.2196/jmir.2642.
7. Ludolph R, Allam A, Schulz PJ. Manipulating Google's Knowledge Graph Box to Counter Biased Information Processing During an Online Search on Vaccination: Application of a Technological Debiasing Strategy. J Med Internet Res. 2016;18(6):137. https://doi.org/10.2196/jmir.5430.
8. Iverson SA, Howard KB, Penney BK. Impact of internet use on health-related behaviors and the patient-physician relationship: a survey-based study and review. J Am Osteopath Assoc. 2008;108(12):699–711.
9. Wald HS, Dube CE, Anthony DC. Untangling the Web-The impact of Internet use on health care and the physician-patient relationship. Patient Educ Couns. 2007;68(3):218–24. https://doi.org/10.1016/j.pec.2007.05.016.
10. Risk A, Dzenowagis J. Review of Internet health information quality initiatives. J Med Internet Res. 2001. https://doi.org/10.2196/jmir.3.4.e28.
11. Viviani M, Pasi G. Credibility in social media: opinions, news, and health information-a survey. Wiley Interdiscip Rev Data Min Knowl Disc. 2017;7(5):1209. https://doi.org/10.1002/widm.1209.
12. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices,. J Epidemiol Community Health. 1999;53(2):105–11. https://doi.org/10.1136/jech.53.2.105.
13. Boyer C, Dolamic L. Feasibility of automated detection of HONcode conformity for health-related websites. Int J Adv Comput Sci Appl. 2014;5(3):. https://doi.org/10.14569/IJACSA.2014.050309.
14. Boyer C, Dolamic L. Automated Detection of HONcode Website Conformity Compared to Manual Detection: An Evaluation,. J Med Internet Res. 2015;17(6):135. https://doi.org/10.2196/jmir.3831.
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. 2017. http://arxiv.org/abs/1706.03762. Accessed 14 Oct 2019.
16. Luong M-T, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. 2015. http://arxiv.org/abs/1508.04025. Accessed 15 Oct 2019.
17. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J. Transformers: State-of-the-art Natural Language Processing. 2019. http://arxiv.org/abs/1910.03771.
18. Ruder S, Peters ME, Swayamdipta S, Wolf T. Transfer Learning in Natural Language Processing. In: Proceedings of the 2019 Conference of the North. Stroudsburg: Association for Computational Linguistics; 2019. p. 15–18. https://doi.org/10.18653/v1/N19-5004.

19. Rees CE, Ford JE, Sheard CE. Evaluating the reliability of DISCERN: A tool for assessing the quality of written patient information on treatment choices. Patient Educ Couns. 2002. https://doi.org/10.1016/S0738-3991(01)00225-7.

20. Khazaal Y, Chatton A, Cochand S, Coquard O, Fernandez S, Khan R, Billieux J, Zullino D. Brief DISCERN, six questions for the evaluation of evidence-based content of health-related websites. Patient Educ Couns. 2009. https://doi.org/10.1016/j.pec.2009.02.016.

21. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. http://arxiv.org/abs/1810.04805.

22. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz682.

23. Allam A, Schulz PJ, Krauthammer M. Toward automated assessment of health Web page quality using the DISCERN instrument. J Am Med Informa Assoc. 2017;24(3):481–87. https://doi.org/10.1093/jamia/ocw140.

24. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

25. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

26. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5(2):157–66. https://doi.org/10.1109/72.279181.

27. Graves A. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, Heidelberg: Springer; 2012. https://doi.org/10.1007/978-3-642-24797-2. http://link.springer.com/10.1007/978-3-642-24797-2.

28. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics; 2014. p. 1724–1734. http://aclweb.org/anthology/D14-1179. Accessed 01 Nov 2019.

29. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 2014. http://arxiv.org/abs/1412.3555. Accessed 01 Nov 2018.

30. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2014. http://arxiv.org/abs/1409.0473. Accessed 18 Dec 2019.

31. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15:1929–58.

32. Bergstra JAMESBERGSTRA J, Yoshua Bengio YOSHUABENGIO U. Random Search for HyperParameter Optimization. J Mach Learn Res. 2012. https://doi.org/10.1162/153244303322533223.

33. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. J Am Med Inform Assoc. 2017;177:. https://doi.org/10.1093/jamia/ocw177.

34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

35. Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel C. MixMatch: A Holistic Approach to Semi-Supervised Learning. 2019. http://arxiv.org/abs/1905.02249. Accessed 18 Dec 2019.

36. Xie Q, Dai Z, Hovy E, Luong M-T, Le QV. Unsupervised Data Augmentation for Consistency Training. 2019. http://arxiv.org/abs/1904.12848.

## Publisher's Note