Journal of Biomedical Informatics 125 (2022) 103976

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Inferring the patient's age from implicit age clues in health forum posts

Christopher M. Black [a], Weilin Meng [a], Lixia Yao [a], Zina Ben Miled [b,c,*]

[a] Merck & Co., Inc., 2000 Galloping Hill Road, Kenilworth, NJ 07033, USA
[b] Department of Electrical and Computer Engineering, IUPUI, Indianapolis, IN 46202, USA
[c] Regenstrief Institute, Inc., 1101 W. 10th Street, Indianapolis, IN 46202, USA

ARTICLE INFO

Keywords:
Patient
Age
Health Forums
Self-Supervised
Classification

ABSTRACT

Broader patient-reported experiences in oncology are largely unknown due to the lack of available information from traditional data sources. Online health community data provide an exploratory way to uncover these experiences at a large scale. Analyzing these data can guide further studies towards understanding patients' needs and experiences. However, analysis of online health data is inherently difficult due to the unstructured nature of these data and the variety of ways information can be expressed over text. Specifically, subscribers may not disclose critical information such as the age of the patient in their posts. In fact, the number of health forum posts that explicitly mention the age of the patient is significantly lower than the number of posts that do not include this information in the Reddit r/Cancer health forum under consideration in the present paper. Health-focused studies often need to consider or control for age as a confounder, hence the importance of having sufficient age data. This paper presents a methodology that can help classify health forum posts according to four age groups (0–17, 18–39, 40–64 and 65 + years) even when the posts do not contain explicit mention of the age of the patient. First, the subset of the posts that include explicit mention of the age of the patient is identified. Second, the explicit age clues are removed from these posts and used to train the proposed age classifier. The resulting classifier is able to infer the age of the patient using only implicit age clues with an average true positive rate (TPR) of 71%. This TPR is comparable to the average TPR of 69% obtained from human annotations for the same set of posts.

1. Introduction

The use of online health forums in health studies offers numerous advantages. These forums often count a large number of subscribers who openly share their experiences with the disease. However, data retrieved from health forums may lack the metadata necessary to the design or interpretation of the results of the health study. The present study was motivated by the need to analyze the side-effects of cancer treatments according to the age of the patient as reported in the online health forum r/Cancer [1].

This use-case is representative of many other health studies such as investigating comorbidity, patients' mental health and quality of life concerns for cancer and for other diseases. In these studies, demographic information, including age, is often used as an inclusion and exclusion criterion when developing the study cohort. It is also used for stratification purposes during analysis. The demographics of patients can have implications on the type of care they would receive, their type of insurance, their disease risks and comorbidities. For instance, age was found to be a significant predictor of the risk for dementia [2] and the frequency of physician visits in older adults [3]. A strong association between adherence to medication and age was also established in hypertensive patients [4]. In addition, incidence and prevalence rates for chronic diseases are often reported according to the age of the patients [5].

The importance of age classification is not limited to health studies. There is a correlation between the age of the author and the age of the consumer in opinion mining [6]. The age of the opinion author is important for several applications in social sciences [7] and for the design of intelligent recommender systems [8].

The data used in this study were extracted from the r/Cancer health forum. This online health forum counts 38,841 current subscribers, and a total of 25 posts and 85 comments over the past 24 h as of June 18, 2021 [9]. Over an extended study period, the number of subscribers and daily posts is lower than these current values for prior months and years. Moreover, some of the posts are non-patient-centric (e.g., recruiting for a clinical trial, advertising a support group, etc.). Patient-centric posts

* Corresponding author at: Department of Electrical and Computer Engineering, IUPUI, 723 W. Michigan St, Indianapolis, IN 46202, USA.
E-mail address: zmiled@iupui.edu (Z. Ben Miled).

https://doi.org/10.1016/j.jbi.2021.103976
Received 25 July 2021; Received in revised form 3 December 2021; Accepted 6 December 2021
Available online 11 December 2021

are the primary focus of this study since they typically describe the patient's experience with cancer whereas the comments are submitted by subscribers in response to these posts. In the patient-centric posts, personal information is disclosed at various levels. For instance, age information is available in less than 30% of the posts extracted from r/Cancer in the present study.

Excluding all posts that do not include explicit mention of the age of the patient will significantly reduce the sample size available to any health study. The aim is to develop a methodology that can classify posts from the online health forum r/Cancer according to the age group of the patient under the following conditions: 1) the target is the age of the patient not that of the subscriber and 2) the age of the patient must be inferred from implicit age clues in the post.

This classification task is a form of second order inference since it depends on implicit age clues in the post. In contrast, a first order inference consists of identifying the age of the patient from posts with explicit age mention. While the underlying methodology was developed primarily for the identification of the age of the patient using implicit age clues, we believe that it can be extended to other NLP applications. For example, second order inference is needed in sarcasm detection [10]. Traditional sentiment analysis relies on explicit clues in the text to classify reviews into negative, positive or neutral. However, sarcasm must be inferred from implicit clues [11]. In the specific case of health studies, second order inference is also needed to estimate disease severity levels [12,13] and assess quality of life measures [14,15].

2. Background

The popularity of online health forums and their availability to researchers led to numerous health studies such as predicting suicide risk levels [16], understanding the impact of the pandemic [17], and identifying emerging drugs involved in overdose [18]. The demographics (e.g., sex, age, race) of the patients are important variables in these studies since they can act as effect modifiers. However, this information is often not available. Earlier attempts at predicting the sex and race of a subscriber to health forums was performed by using rule-based classification derived from aggregated census data [19].

In general, previous related studies can be organized under two main categories: author profiling and demographic traits identification. An example of recent work from each of these categories is reviewed next. Author profiling consists of classifying the author into different classes primarily relying on the writing style. In [20], an author profiling model was developed and tested on comments submitted by students participating in several online courses. The model consists of an ensemble of three complimentary components that use a different level of text representation each: character, word and phrase. This model was able to classify the students according to three categories: working, not working, and retired with a high level of accuracy.

The focus of demographic traits identification is to classify sequences of text according to the presence or absence of identifying traits (e.g., occupation, name, age, etc.). This information can assist, for instance, in the de-identification of medical notes as investigated in [21]. In this study, different models including BERT, Naive Bayes, CNN, LSTM and Random Forest were compared. The BERT model was found to have the highest precision and recall in identifying sentences containing demographic traits. A similar BERT-based model is used in the present paper. However, the objective in the present paper is different. Our aim is to establish the value of a specific demographic trait, namely age. Extracting structured facts from the demographic traits was left for future work in [21].

Few earlier studies focused on predicting the age of the patient from other sources of data such as predicting the skeletal age from hand radiographs [22] and predicting the chronological age from raw MRI data [23]. In other domains, research on estimating the age of a person from facial images is extensive and typically motivated by applications such as security identification and forensic art [24,25].

The methodology proposed in the present paper starts with posts that explicitly mention the age of the patient, removes these explicit age clues from the posts and subsequently trains a model to estimate the age of the patient from the remaining implicit age clues in the post. In the literature, the removal of the explicit clues is often referred to as "masking". Training a machine learning model using masked data is not new. The use of masking can be traced back to at least the continuous bag of word (CBOW) model [26]. CBOW is a neural network model that is trained to predict the missing word from the surrounding contextual words. The output of the hidden layer nodes is then extracted and used to represent the word. These word vector representations are referred to word embedding in the literature. CBOW embeddings were shown to be useful in numerous NLP applications [27–29]. This technique was further developed by several researchers and led to a family of pre-trained language models including the Bidirectional Encoder Representations from Transformers (BERT) [30] and Longformer [31] among many others. The BERT language model introduced bidirectional (left and right) context in all layers of the neural network making it possible to develop task-specific models with limited fine tuning. However, this contextual information is captured with an attention mechanism that scales quadratically with the length of the input sequence. Thus, it is difficult for BERT to process long sequences of text. Longformer modifies this architecture by using an attention layer that scales linearly with the length of the input sequence. Instead of using a global attention layer for all tokens in the text sequence, Longformer uses the global attention layer for only a few pre-selected positions in the text. It then complements this global attention layer with a local attention layer that is limited in context by a dilated sliding window around the target tokens. The ability of Longformer to process long input sequences made it the model of choice for the present study since most health forum posts are long.

While in principle, masking refers to the removal of the target outcome, the masking operation adopted in this paper is different from the specific word removal in CBOW. In the present paper, one or more phrases that include explicit mention of the age of the patient are removed. Moreover, the objective of the model is to predict the age of the patient rather than generate a vector representation for a given word. That said, the use of masking in the present paper is inspired by CBOW and other previous research studies. Masking is a form of self-supervised learning which is also used in image [32] and signal [33,34] processing applications. For these applications, pretext tasks are defined on labels that can be directly generated from the data. A pretext task is, for example, associating a rotated image with the original one. The model is trained on these pretext tasks in the hope that it will also learn to make higher order inferences.

For NLP applications, self-supervised learning has been primarily used for sentiment classification [35]. Self-supervised in this case often implies a hybrid learning approach that combines a lexicon-based classifier with a supervised machine learning classifier [36,37]. For instance, a lexicon-based classifier can be trained to generate a positive or negative label as well as a confidence score for a sample text. The samples with the highest confidence score from this first classifier along with their predicted labels are then used to train the second classifier [37].

Self-supervised learning in the present paper refers to the use of masked text data and corresponding labels to train the model to infer the age of the patient from posts with only implicit age clues. As such, the proposed approach is closer to training with pretext tasks in image and signal processing applications.

3. Methods

The data used in this study were extracted from a public archive [38] of the r/Cancer health forum from the period starting 1/1/2014 to 30/4/2020. Only posts are considered in this study. Comments in the form of replies to the original posts are excluded. Moreover, posts that do not

focus on a patient (e.g., posts recruiting for a clinical trial) were also excluded. These are typically identified by a hypertext metadata. All posts are anonymous and consist of a title followed by a message body.

The aim is to assign each post to one of the four age groups: 0–17 years, 18–39 years, 40–64 years, and 65 + years in accordance with the current age of the patient or the age of the patient at the time of his/her death, as opposed to, for example, the age of the patient at the time of diagnosis. The above age groups correspond to childhood, early, middle, and older adulthood, respectively. Different types of age stratification are used in the literature. For example, the US census stratifies the age of the general population over consecutive periods of 5 years [39]. When reporting cancer incidence and prevalence statistics, the CDC stratifies the age of the patients over consecutive periods of 10 years [5]. Cancer prevention and risk literature by the CDC is stratified according to the four primary age groups used in this study [40]. The granularity of the age group is often dictated by the nature of the data. Ideally, the exact age of the patient or age groups with fine granularity are preferred. Given that the data was collected from an online health forum, as opposed to using traditional health surveys, a finer age granularity would have resulted in the ambiguous classification of several posts (e.g., *5 years ago …in her fifties*). Moreover, several previous cancer studies [41,42] use the age stratification adopted in the present paper as it aligns with the CDC cancer prevention and risk age groups [40].

The example posts in Table 1 include different types of age clues: explicit/direct, explicit/indirect and implicit. These posts can be submitted by either the patients or the caregivers. In the first post (p_1), the current age of the patient is explicitly mentioned (i.e., *35 years old*) and this post would be assigned to the age group 18–39 years. This an instance of an explicit and direct mention of the current age of the patient. This post is also an example of a post where the subscriber is a caregiver and not the patient. In the second post (p_2), the current age of the patient is also explicitly mentioned (i.e., *3 years ago when I was 15*). However, this is an indirect mention of the current age of the patient. This post will also be assigned to the age group 18–39 and is an example of the case where the subscriber is the patient. In the third post (p_3), there is no explicit reference to the current age of the patient. Nonetheless, there are two clues that may help infer the age group of the patient: The patient was diagnosed at the age of 17; and the patient is currently in college. Using these two clues, the current age group of the patient may be estimated to be 18–39 years. This inference is difficult, not with absolute certainty and shows an example of the age inference from implicit age clues that is being pursued in this study.

Table 1
Example posts with explicit (italics) and implicit (underline) age clues. Explicit age clues can either directly or indirectly mention the age of the patient.

Explicit-direct age information (p_1)
Ideas for gifts, She is still alive On March it will be our 5th anniversary, she recently was told that her cancer has spread to her pelvis so she is having problems to walk, and currently there is no medical treatment for her. But hey! she is still alive and I want to give her something special, it could be our last anniversary. She is *35 years old* any ideas for a gift?
Explicit-indirect age information (p_2)
Different stages of cancer, what do they mean? I was diagnosed with synovial sarcoma *3 years ago when I was 15*, but I was never told what stage I was in or anything like that. The doctors said that I had the tumor in my foot for approximately 5 years, does this have anything to do with the stage of cancer? Every time I read about someone that has had cancer, they always talk about what stage they are in but I don't really understand what it means.
Implicit age information (p_3)
How long did your chemo brain last? I had a germ cell tumor <u>when I was 17</u> and received three rounds of BEP. I obviously had issues with Brain fog, but I thought it faded away. Well, <u>Im in college now</u> and have the hardest time concentrating and paying attention. Wondering if it could still be the brain fog?? Curious about other peoples experiences. Thanks:)

3.1. Annotation

Each post was annotated by three annotators. The annotators were asked to review each post, identify explicit age clues, if any, and assign the post to one of the age groups. An annotation protocol was developed through an iterative process that ensured high quality annotation. The annotators were also trained on sample posts from each category. The quality of the annotation was then reviewed by an independent annotator. Annotation for NLP tasks is time consuming. That said, high quality annotation is necessary to develop an accurate language model. Fortunately, a limited set of highly curated posts is often sufficient.

In the first round, the annotators had to assign each post to an age group only when there is an explicit mention of the age of the patient in the text. Moreover, they had to provide the starting and ending position of all age clues by highlighting them in the annotation application. This annotation step was not only useful in enhancing the quality of the annotation but also made it possible to remove the explicit age clues from the post. Posts without explicit age clues are assigned to the unknown age group. In addition, the annotators were asked to identify whether the subscriber is the patient or the caregiver. This information is used to understand the distribution of the posts and to estimate the number of posts that would have been excluded if a health study was limited to posts where the subscriber is the patient.

This annotation process generated the following two disjoint set of posts:

- $E = \{(p_i, c_i, y_i)\}_{i=1}^{N}$ where p_i is the text of the post, c_i is the set of explicit age clues, y_i is the age group of the patient, and N is the total number of posts. The example post (p_1, {*35 years old*}, 18–39) from Table 1 belongs to this set.
- $I = \{(p_j, \varnothing, \varnothing)\}_{j=1}^{L}$ where p_j is the text of the post and L is the total number of posts. These posts do not include any explicit age clues and the age group of the patient is unknown. The example post (p_3,∅, ∅) from Table 1 belongs to this set.

A post may include several explicit/direct age clues. Sometimes, the same age clue is repeated multiple times, typically, in the title, the body and sometime at the end of the post as demarked by the hashtag TLDR. The age clue can also appear in different forms (e.g., *"I'm only 21"* and later in the text *"in my twenties"*). Moreover, the age of the patient can be derived indirectly from multiple explicit/indirect age clues as in the case of p_2 in Table 1. These clues may be separated by multiple sentences (e.g., *"…30 years cancer survivor…One kidney was removed when I was 4…"*). All explicit age clues in the post, whether direct or indirect, are collected by the annotators.

Table 2 shows the result of the masking operation when applied to the first two posts of Table 1. This operation eliminates the explicit age clues from the posts including all the clues that directly mention the current age of the patient (e.g., $p1$). When the age of the patient can be derived from multiple clues (i.e., indirect mention of the current age of

Table 2
Masking of the example posts from Table 1 to remove explicit age clues. The only remaining age clues are implicit as underlined in the text.

$\widehat{p_1}$	Ideas for gifts, She is still alive On March it will be our <u>5th anniversary</u>, she recently was told that her cancer has spread to her pelvis so she is having problems to walk, and currently there is no medical treatment for her. But hey! she is still alive and I want to give her something special, it could be our last anniversary. She is <*omitted text*> any ideas for a gift?
$\widehat{p_2}$	Different stages of cancer, what do they mean? I was diagnosed with synovial sarcoma <u>3 years ago</u> when <*omitted text*>, but I was never told what stage I was in or anything like that. The doctors said that <u>I had the tumor in my foot for approximately 5 years</u>, does this have anything to do with the stage of cancer? Every time I read about someone that has had cancer, they always talk about what stage they are in but I don't really understand what it means.

the patient), only the reference age is removed. For example, when masking is applied to $p2$, the clue *"I was 15"* is removed while *"3 years ago"* is retained. The goal of the masking operation is to translate posts with explicit age clues to posts with only implicit age clues. Typically, masking involves replacing the original text with a special token (e.g. *MASK*). However, in the present paper, the age clue is simply removed from the text to 1) avoid any position or context-related information leak and 2) mimic posts with only implicit age clues which would not include the special token. The masking operation as described above is denoted by $\widehat{(\cdot)}$ in the remainder of the paper. The new set of posts $\widehat{p_i}$ resulting from the masking of the posts p_i in E is defined as follows:

- $\widehat{E} = \{(\widehat{p_i}, \varnothing, y_i)\}_{i=1}^{N}$. Even though explicit age clues were removed from the original posts (p_i) to generate the modified posts ($\widehat{p_i}$), the age labels y_i are still available because they were obtained prior to masking.

In a second round, the annotators were asked to label randomly sampled posts from $\widehat{E}$ and I. These samples do not include any explicit age clues. For the former set of posts, this is by design (i.e., due to masking). For the latter, these posts did not originally include any explicit age clues as identified by the annotators in the first round. Three annotations were also collected for these posts. However, in this second round the annotators were asked to provide a best-guess estimate as to the age of the patient since the posts do not include explicit age clues. They had to infer the age of the patient using implicit clues such as kinship terms (e.g., grandfather), education level (e.g., in college), professional level (e.g., retired), or daily activities (e.g., going to the gym). Inferring the age group of the patient for these posts is subjective.

3.2. Age Classifier

Two classification models were developed. They are both structured as multi-class classifiers where each age group is a nominal label. The first model is trained with a subset of the posts from E. This model is denoted M_E. The second model is trained with masked posts from $\widehat{E}$ and is denoted $M_{\widehat{E}}$. Both models are trained with ground truth age labels which are extracted from the explicit age clues in the original posts.

The distribution of the posts in E and $\widehat{E}$ across the age groups is not uniform. Age groups 0–17 and 65 + are the minority classes and age groups 18–39 and 40–65 are the majority classes. All of the posts in the minority classes are retained. The majority classes are down-sampled in order to avoid class imbalance. The posts are then randomly split 80/20 in each age group for training and testing purposes. The split resulted into two subsets E^{tr} and E^{ts} from E. Similarly, two subsets are also generated from the masked set $\widehat{E}$: $\widehat{E}^{tr}$ and $\widehat{E}^{ts}$.

The classifiers M_E and $M_{\widehat{E}}$ are trained using Longformer [31]. As mentioned above, this language model was selected because it can accommodate long text. Posts from the r/Cancer health forum tend to exceed the limits imposed by other language models [30]. The training was performed using the AdamW optimizer and the default hyperparameters values provided in [31]. The batch size and the learning rate were set to 16 and $2e^{-5}$, respectively. Tuning was not performed on these hyperparameters. Model selection with respect to the number of epochs was achieved using nested 5-fold cross-validation on the training dataset. The number of epochs was varied from 2 to 10. For each cross-validation round, the model was trained with 4 out of the 5 groups from the training dataset and evaluated on the held-out group. No significant improvements were observed in the average and standard deviation of the accuracy of both classifiers across the 5 folds beyond 5 epochs. Therefore, the final models (i.e., M_E and $M_{\widehat{E}}$) were both trained using the entire training dataset and 5 epochs.

M_E is trained with posts in E^{tr} and $M_{\widehat{E}}$ is trained with posts in $\widehat{E}^{tr}$. While the two models were developed using an off-the-shelve pretrained language model, the present study shows the importance of the data used to train these two models. M_E is trained with posts with explicit age mention and therefore struggles with inferring the age of the patient from posts with only implicit age clues. The second model, $M_{\widehat{E}}$, is trained with posts that only include implicit age clues and therefore developed the ability to infer the age group of the patient even in the absence of explicit age clues. To demonstrate this difference in inference level, the two models are tested on posts with explicit age mention from E^{ts} and posts without explicit age mention from $\widehat{E}^{ts}$. The age labels generated by the two models are compared to the ground truth labels using the true positive rate (TPR) for each age group. In addition, the results of the models are also compared to the labels generated by three annotators for the posts in $\widehat{E}^{ts}$. The purpose of this latter analysis is to evaluate the accuracy of the models compared to human annotation when inferring the age group of the patient from posts without explicit age clues.

Finally, since the aim of the study is to estimate the age group of the patient from posts which originally contain only implicit age clues, the models M_E and $M_{\widehat{E}}$ were also tested on randomly selected posts from the set I. Posts in this set were identified as "patient's age unknown" in the first annotation round and therefore do not include any explicit age clues. These posts were also annotated by three annotators who provided their best *estimate* for the age of the patient. The comparison between the classification results from the models and the labels generated by the annotators was performed by using the parity metric since the ground truth labels are not available in this case. Parity is the percent agreement for positive labels in each age group.

4. Results

The total number of posts in this study is 16,768. These posts are divided into two sets. The set E consists of 4,738 posts (28%) with explicit age clues. The set I consists of the remaining 12,030 (72%) posts with only implicit age clues. The distribution of E according to the age group of the patient and to whether the subscriber is the patient or the caregiver is provided in Table 3. For I, the age group distribution is not available since the posts do not mention the exact age of the patient. The patient/caregiver split in I is 42%/58%.

The distribution of the posts in E across the age groups is non-uniform. Age groups 0–17 and 65 + are the minority classes and age groups 18–39 and 40–65 are the majority classes. In order to promote class balance, a total of 600 randomly selected posts from the majority age groups are retained. The reduced set of posts includes all 361 posts from the age group 0–17, all 557 posts from the age group 65 + and the 600 posts resulting from the down sampling of each of the majority age groups 18–39 and 40–64 for a total of 2,118 posts. These posts are then split 80/20 along each age group for training (E^{tr}) and testing (E^{ts}). The number of posts in E^{ts} is provided in the first column of Table 4.

Masking was then applied to E^{ts} and E^{tr} to remove the explicit age clues from the posts. The resulting masked sets of posts are labeled $\widehat{E}^{ts}$ and $\widehat{E}^{tr}$, respectively. The first model, M_E, is trained with the set E^{tr} and therefore has access to explicit age clues in the posts. The second model, $M_{\widehat{E}}$, is trained with $\widehat{E}^{tr}$ and is not exposed to explicit age information.

Table 4 shows the TPR values for M_E and $M_{\widehat{E}}$ when tested on E^{ts} (with

Table 3
Distribution of the posts in E according to the age group of the patient and whether the subscriber is the patient or the caregiver.

Age group	Patient	Caregiver	Total
0–17	148 (41%)	213 (59%)	361 (8%)
18–39	1881 (74%)	647 (26%)	2,528 (53%)
40–64	196 (15%)	1,096 (85%)	1292 (27%)
65+	11 (2%)	546 (98%)	557 (12%)

Table 4

True positive rate (TPR) for M_E,$M_{\widehat{E}}$, and annotators for the test posts in E^{ts} (with explicit age clues) and in $\widehat{E}^{ts}$ (without explicit age clues). [1] Δ is the difference between the TPR obtained for posts in E^{ts} and $\widehat{E}^{ts}$ for each model. [2] TPR from the consensus among the three annotators when labeling the posts in $\widehat{E}^{ts}$. The p-values are calculated for the labels generated by the annotators for posts in each age group by using the Friedman test.

Age	Number	M_E			$M_{\widehat{E}}$			Annotators[2]
Group	of posts	E^{ts}	$\widehat{E}^{ts}$	Δ[1]	E^{ts}	$\widehat{E}^{ts}$	Δ[1]	$\widehat{E}^{ts}$ (p-value)
0–17	74	93%	57%	36%	89%	78%	11%	65% (0.236)
18–39	120	98%	83%	15%	89%	86%	3%	88% (0.634)
40–64	120	98%	38%	60%	43%	48%	−5%	68% (0.914)
65+	113	99%	64%	35%	83%	70%	13%	53% (0.099)
Average	427	97%	60%	37%	76%	71%	5%	69%

explicit age clues) and $\widehat{E}^{ts}$ (without explicit age clues). The table also shows the TPR values for the labels obtained from the three annotators after consensus for the masked posts in $\widehat{E}^{ts}$. When labeling $\widehat{E}^{ts}$, the annotators can assign different age groups to a given post. The p-value values in Table 4 indicate that this difference is not statistically significant. The last row of the table is the average TPR for all the posts in each set. It should be noted that the posts in E^{ts} and $\widehat{E}^{ts}$ are the same, with the exception that explicit age clues are removed (i.e., masked) in the case of $\widehat{E}^{ts}$ as exemplified by the transformation $p_1 \rightarrow \widehat{p}_1$ and $p_2 \rightarrow \widehat{p}_2$ in Tables 1 and 2.

Several observations can be made from the results in Table 4. The model M_E has a high TPR (i.e., ⩾93%) when tested with posts that include explicit age clues (E^{ts}). This model was trained with posts that include explicit age clues (E^{tr}) and the results indicate that M_E is able to accurately infer the age of the patient given these explicit age clues. This case also establishes a reference baseline and as such, any deviation thereafter can only be attributed to the training or testing data.

When M_E is tested with posts from $\widehat{E}^{ts}$, the results do not show the same level of accuracy. In fact, the average TPR drops from 97% for E^{ts} down to 60% for $\widehat{E}^{ts}$. We can conclude from this drop that M_E is unable to fully utilize implicit age clues in the posts. Moreover, this drop in performance is not the same across the age groups. For instance, M_E is still able to infer the age of the patients in the age group 18–39 from masked posts with a TPR of 83% despite the absence of explicit age clues. The largest drop in TPR between the test sets E^{ts} and $\widehat{E}^{ts}$ for M_E is observed for the age group 40–64 (Δ = 98% - 38% = 60%). The actual mis-classifications for this model with both test datasets are provided in Appendix A. For E^{ts}, the mis-classifications are all within the adjacent age groups whereas for $\widehat{E}^{ts}$ they extend to non-adjacent age groups.

The model $M_{\widehat{E}}$ shows lower TPR values than those of M_E for the set of posts in E^{ts}. This is expected since $M_{\widehat{E}}$ was not trained with posts that include explicit age clues. Thus, it is unable to leverage this information. The TPR values for $M_{\widehat{E}}$ are still higher than 83% for all age groups expect for the age group 40–64. The TPR values for $M_{\widehat{E}}$ are, however, consistently higher than those of M_E when tested on $\widehat{E}^{ts}$. The average TPR values produced by $M_{\widehat{E}}$ and M_E for this masked set of posts are 71% and 61%, respectively. Moreover, as shown by the difference columns (Δ) in Table 4, $M_{\widehat{E}}$ does not suffer as much as M_E from the removal of the explicit age clues. Actually, a 5% improvement in TPR for the age group 40–64 is observed for $M_{\widehat{E}}$ between E^{ts} (with explicit age clues) and $\widehat{E}^{ts}$ (without explicit age clues). The confusion matrices for $M_{\widehat{E}}$ when tested on E^{ts} and $\widehat{E}^{ts}$ are provided in Appendix A. Compared to M_E, the mis-classifications of $M_{\widehat{E}}$ when tested on $\widehat{E}^{ts}$ are more constricted to adjacent age groups.

On average, $M_{\widehat{E}}$ performs better and M_E performs worst than the annotators on the masked posts in $\widehat{E}^{ts}$. In increasing order, the average TPR values are 60%, 69% and 71% for M_E, annotators, and $M_{\widehat{E}}$, respectively. Specifically, $M_{\widehat{E}}$ has higher TPR for the age groups 0–17 and 65 + compared to the annotators, similar TPR for the age group 18–39 and lower TPR for the age group 40–64.

The above results evaluate the ability of the models and the annotators to infer the age of the patient from posts with explicit age clues and from posts where these clues have been removed. The key question is how would these results translate to posts which originally did not include any explicit age clues. Since ground truth age labels are not available in this case, the evaluation can only be performed by using the pairwise agreement (parity) in label assignments between $M_{\widehat{E}}$ and M_E - PR($M_{\widehat{E}}$,M_E) - and between $M_{\widehat{E}}$ and the annotators - PR($M_{\widehat{E}}$, Annotators). These parities for the 600 posts randomly sampled from I vary by age group as shown in Table 5.

The agreement between the labels generated by $M_{\widehat{E}}$ and those generated by either M_E or the annotators for the age group 0–17 years is low. According to Table 4, both M_E and the annotators show low TPR values for the masked posts in this age group which may explain the observed low parity.

For the age group 18–39, the pairwise agreement between $M_{\widehat{E}}$ and either M_E or the annotators is high (⩾88%). This result aligns with the high TPR values for the masked posts in $\widehat{E}^{ts}$ (Table 4). The parity between $M_{\widehat{E}}$ and the annotators (71%) is higher than that between $M_{\widehat{E}}$ and M_E (50%) for the age group 40–64. For the fourth age group, 65+, PR($M_{\widehat{E}}$, M_E) is 85% and PR($M_{\widehat{E}}$, Annotators) is 33%. This parity is consistent with the reported TPR in Table 4 for posts in $\widehat{E}^{ts}$ from this age group. In fact, the lowest TPR by the annotators was reported for this age group.

5. Discussion

Demographic information is necessary for our understanding of population behavior and trends. Unfortunately, this information is hard to collect and not always readily available [8,19]. Estimating age, one of the key demographic, was thoroughly investigated in the case of image data for both health and security purposes [22–25]. Estimating age from text data received limited attention.

The goal of the present study is to extract the age of the patient from posts that are submitted to online health forums when the age of the patient is not explicitly mentioned in the text. It is possible to approach this task from the perspective of author profiling. However, when the subscriber is a caregiver, author profiling and writing styles do not

Table 5

Parity of the model $M_{\widehat{E}}$ compared to the labels generated by M_E and the annotators for 600 posts randomly sampled from I (implicit age clues).

Age group	Number of posts	PR($M_{\widehat{E}}$,M_E)	PR($M_{\widehat{E}}$, Annotators)
0–17	56	34%	18%
18–39	265	89%	88%
40–64	157	50%	71%
65+	122	85%	33%

provide useful information. Author profiling can only help when the subscriber is the patient [20]. An alternative is to approach the task from the perspective of demographic traits identification while specifically focusing on a single trait, age. Demographic traits identification was explored for the purpose of de-identifying medical notes [21]. Since medical notes are authored by health care providers, this scenario is similar to the case where the subscriber is a caregiver. In this previous study, text sequences were classified as containing identifying traits but the values of the traits were not established [21]. As suggested by the authors, extracting the value of a given trait requires a specific classifier for each trait. The current study demonstrates that this approach is possible without the need for distinguishing between authors that are patients or caregivers and even when only implicit age information is included in the text.

Compared to other traits, extracting values for the age trait is hard because of the multi-class nature of this task. Extracting the value of the working status was recently demonstrated in [20]. Classifiers were also developed to extract the values of sex and race from online posts [19]. Similarly, a statistical model for race, sex and income was proposed for proportion distribution with respect to political opinion mining from Twitter data [7]. However, none of these previous studies investigated the age of the author.

The present paper outlines a methodology that estimates the age of the patient from health forum posts. This use-case was primarily motivated by the desire to extend health studies beyond traditional sources of data such as surveys and electronic medical records. Health forums are an emerging source of information that can provide extensive insight into the patient's experience with various disease conditions. Unfortunately, they are also an example of non-traditional data collections where age information may not be readily available.

In fact, despite the extended data collection period (i.e., approximately 6.5 years) in the present study, Table 3 indicates that limiting the study to posts where 1) the age of the patient is explicitly mentioned and 2) the patient is also the subscriber, is equivalent to an attrition from 16,768 original posts to only 1,536 posts that satisfy the above criteria. Using the proposed methodology allows the inclusion of additional posts from the caregiver and the unknown age group categories with an accuracy comparable to human annotation. This methodology is based on an age classifier $M_{\widehat{E}}$ which is trained on posts without explicit age clues. As a result, $M_{\widehat{E}}$ is compelled to estimate the age of the patient using only implicit age clues. On average, this estimate is better than the age estimate produced by M_E which was trained with posts that include explicit age clues and better than the age estimate derived from the consensus among three annotators (Table 4).

The inference ability of $M_{\widehat{E}}$ varies among the age groups. Some age groups are easier to predict than others with either explicit or implicit age clues. For example, the age group 18–39 consistently shows the highest TPR values (⩾83%) for posts with explicit age clues (E^{ts}) and without age clues ($\widehat{E}^{ts}$) for the models $M_{\widehat{E}}$ and M_E as well as the annotators. The pairwise agreements for posts with implicit age clues (I) between $M_{\widehat{E}}$ and M_E and between $M_{\widehat{E}}$ and the annotators are also the highest (⩾88%) for this age group.

In contrast, the age group 40–64 is the hardest to predict for both $M_{\widehat{E}}$ and M_E. The best TPR for this age group was obtained by the annotators for the masked data set $\widehat{E}^{ts}$. The parity between $M_{\widehat{E}}$ and the annotators for the posts from I in this age group is 71% (Table 5). This indicates that despite the low TPR associated with this age group, there is still a good agreement between the labels generated by $M_{\widehat{E}}$ and the annotators.

For the age groups 0–17 and 65+, $M_{\widehat{E}}$ is better than both the annotators and M_E at estimating the age of the patient from posts without explicit age clues. The TPR values produced by $M_{\widehat{E}}$ are higher than those of the annotators for the two age groups by 13% and 17% (Table 4), respectively. Similarly, these values are higher than those of M_E by 21% and 6%, respectively.

The above results show that it is possible to develop an age classifier that can deliver annotation on par with human annotation for posts with implicit age clues. Notwithstanding, the fact that the annotators are able to exploit the ordinal nature of the age groups whereas $M_{\widehat{E}}$ was trained using nominal labels. The results also show that there is a major difference between first order and second order inferences in NLP classification applications.

That said, there are several aspects of the proposed methodology that need to be taken into consideration. The age distribution of the cancer patients in E (Table 3), which is limited to the posts with explicit mention of the age of the patient, may not reflect the age distribution of the entire set of posts in r/Cancer which consists of posts from both E and I. Moreover, the age distribution in E may also not reflect the age distribution of cancer patients in the general population. The low percentage of patients in the age group 0–17 is probably due to the low incidence rate of the disease for this age group in the general population [5]. However, the low count of patients observed in E for the 65+ age group does not correspond to the incidence rate in the general population [5]. This is probably due to the lack of representation of this age group in the active members of social media [43,25].

Extending the proposed inference methodology from implicit clues to other types of demographics also has limitations. Estimating the sex of the patient is only interesting for posts that are submitted by the patients themselves. Indeed, these posts include limited sex clues. In contrast, posts submitted by caregivers often include several sex clues in the form of pronouns (e.g., he/she) or kinship terms (e.g., mother/father, sister/brother). Masking sex clues can also be more difficult if they occur frequently in the text.

In the case of race, the limitation is of a different type. Based on the review of the r/Cancer posts in this study, there is a limited number of posts that actually mention the race of the patient. Therefore, the data needed to train the model may be insufficient. However, there are other health variables for which the proposed methodology may be applicable including cancer staging, cancer type or disease severity. Extending the proposed approach to these variables is the subject of future work.

6. Conclusion

Being able to ascertain the current age of the patient when there is no explicit age clues in the health forum post is important for many health studies. The age of the patient can be a confounding or a control variable. This paper presents a methodology that can help estimate the age of the patient from these posts. For posts where the explicit age clues have been intentionally removed, both the proposed model ($M_{\widehat{E}}$) and the human annotators achieved comparable TPR values. For posts which originally included only implicit age clues, the proposed model and the annotators have high parity for age groups 18–39 and 40–64. These results suggest that the methodology can be used to alleviate the tedious annotation effort needed by various health studies.

The models developed in this study and related documentation are available at *osf.io/4xawc*.

Authors contributions

All authors equally contributed to the development of the design methodology, implementation of the models and analysis of the results. All authors reviewed and approve the current manuscript.

Role of Funding Source

Authors are responsible for the work described in this paper. Study sponsor had no involvement in the collection, analysis and interpretation of the data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported in part by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA. The authors would like to thank Jarod Baker of the Regenstrief Institute for his support. Special thanks are also extended to the reviewers for their valuable comments.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2021.103976.

References

[1] Cancer: Discussion & Support, https://www.reddit.com/r/cancer/ (accessed December 15, 2020).
[2] Z. Ben-Miled, K. Haas, C.M. Black, R.K. Khandker, V. Chandrasekaran, R. Lipton, M.A. Boustani, Predicting dementia with routine care emr data, Artif. Intell. Med. 102 (2020) 101771, https://doi.org/10.1016/j.artmed.2019.101771.
[3] T. Hu, N.D. Dattani, K.A. Cox, B. Au, L. Xu, D. Melady, L. Jaakkimainen, R. Jain, J. Charles, Effect of comorbidities and medications on frequency of primary care visits among older patients, Can. Fam. Physician 63 (1) (2017) 45–50.
[4] S.J. Kim, O.D. Kwon, E.B. Han, C.M. Lee, S.-W. Oh, H.-K. Joh, B. Oh, H. Kwon, B. Cho, H.C. Choi, Impact of number of medications and age on adherence to antihypertensive medications: a nationwide population-based study, Medicine 98 (49). doi:10.1097/MD.0000000000017825.
[5] U.C.S.W. Group, U.S. Cancer Statistics Data Visualizations Tool, based on 2019 submission data (1999–2017): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, www.cdc.gov/cancer/dataviz (accessed February 2, 2021).
[6] J.A. Balazs, J.D. Velásquez, Opinion mining and information fusion: a survey, Information Fusion 27 (2016) 95–110, https://doi.org/10.1016/j.inffus.2015.06.002.
[7] E.M. Ardehaly, A. Culotta, Mining the demographics of political sentiment from twitter using learning from label proportions, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 733–738. doi:10.1109/ICDM.2017.84.
[8] J. Beel, S. Langer, A. Nürnberger, M. Genzmehr, The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2013, pp. 396–400. doi:10.1007/978-3-642-40501-3_45.
[9] r/cancer stats, https://subredditstats.com/r/cancer (accessed June, 2021).
[10] A. Joshi, P. Bhattacharyya, M.J. Carman, Automatic sarcasm detection: A survey, ACM Computing Surveys (CSUR) 50 (5) (2017) 1–22, https://doi.org/10.1145/3124420.
[11] A. Kumar, V.T. Narapareddy, V.A. Srikanth, A. Malapati, L.B.M. Neti, Sarcasm detection using multi-head attention based bidirectional lstm, Ieee Access 8 (2020) 6388–6397, https://doi.org/10.1109/ACCESS.2019.2963630.
[12] B. Gallo Marin, G. Aghagoli, K. Lavine, L. Yang, E.J. Siff, S.S. Chiang, T.P. Salazar-Mather, L. Dumenco, M.C. Savaria, S.N. Aung, et al., Predictors of covid-19 severity: A literature review, Reviews in medical virology 31 (1) (2021) 1–10, https://doi.org/10.1007/s11606-020-05889-w.
[13] S.K. Nutley, A.M. Falise, R. Henderson, V. Apostolou, C.A. Mathews, C.W. Striley, Impact of the covid-19 pandemic on disordered eating behavior: Qualitative analysis of social media posts, JMIR mental health 8 (1) (2021) e26011, https://doi.org/10.2196/26011.
[14] M. De Choudhury, S. De, Mental health discourse on reddit: Self-disclosure, social support, and anonymity, in: Eighth international AAAI conference on weblogs and social media, 2014.
[15] M.M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum, IEEE Access 7 (2019) 44883–44893, https://doi.org/10.1109/ACCESS.2019.2909180.
[16] V. Ruiz, L. Shi, W. Quan, N. Ryan, C. Biernesser, D. Brent, R. Tsui, Clpsych2019 shared task: Predicting suicide risk level from reddit posts on multiple forums, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 162–166. doi:10.18653/v1/W19-3020.
[17] D.M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, S.S. Ghosh, Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study, Journal of medical Internet research 22 (10) (2020) e22635, https://doi.org/10.2196/22635.
[18] A.P. Wright, C.M. Jones, D.H. Chau, R.M. Gladden, S.A. Sumner, Detection of emerging drugs involved in overdose via diachronic word embeddings of substances discussed on social media, J. Biomed. Inform. (2021) 103824, https://doi.org/10.1016/j.jbi.2021.103824.
[19] S.A. Sadah, M. Shahbazi, M.T. Wiley, V. Hristidis, A study of the demographics of web-based health-related social media users, Journal of medical Internet research 17 (8) (2015) e194, https://doi.org/10.2196/jmir.4308.
[20] T. Aljohani, A.I. Cristea, Predicting learners' demographics characteristics: Deep learning ensemble architecture for learners' characteristics prediction in moocs, in: Proceedings of the 2019 4th International Conference on Information and Education Innovations, 2019, pp. 23–27. doi:10.1145/3345094.3345119.
[21] A. Feder, D. Vainstein, R. Rosenfeld, T. Hartman, A. Hassidim, Y. Matias, Active deep learning to detect demographic traits in free-form clinical notes, J. Biomed. Inform. 107 (2020) 103436, https://doi.org/10.1016/j.jbi.2020.103436.
[22] S.S. Halabi, L.M. Prevedello, J. Kalpathy-Cramer, A.B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L.A. Pereira, R.T. Sousa, N. Abdala, et al., The rsna pediatric bone age machine learning challenge, Radiology 290 (2) (2019) 498–503, https://doi.org/10.1148/radiol.2018180736.
[23] J.H. Cole, R.P. Poudel, D. Tsagkrasoulis, M.W. Caan, C. Steves, T.D. Spector, G. Montana, Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker, NeuroImage 163 (2017) 115–124, https://doi.org/10.1016/j.neuroimage.2017.07.059.
[24] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: A survey, IEEE transactions on pattern analysis and machine intelligence 32 (11) (2010) 1955–1976, https://doi.org/10.1109/TPAMI.2010.36.
[25] A. Clapés, O. Bilici, D. Temirova, E. Avots, G. Anbarjafari, S. Escalera, From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2373–2382. doi:10.1109/CVPRW.2018.00314.
[26] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
[27] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, H. Liu, A comparison of word embeddings for the biomedical natural language processing, Journal of biomedical informatics 87 (2018) 12–20, https://doi.org/10.1016/j.jbi.2018.09.008.
[28] L. Duong, H. Kanayama, T. Ma, S. Bird, T. Cohn, Multilingual training of crosslingual word embeddings, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 894–904. doi:10.18653/V1/E17-1084.
[29] A.B. Dieng, F.J. Ruiz, D.M. Blei, Topic modeling in embedding spaces, Transactions of the Association for, Computational Linguistics 8 (2020) 439–453, https://doi.org/10.1162/tacl_a_00325.
[30] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019. arXiv:1810.04805, doi:10.18653/v1/N19-1423.
[31] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150.
[32] A. Kolesnikov, X. Zhai, L. Beyer, Revisiting self-supervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1920–1929. doi:10.1109/CVPR.2019.00202.
[33] P. Sarkar, A. Etemad, Self-supervised learning for ecg-based emotion recognition, in: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 3217–3221. doi:10.1109/ICASSP40776.2020.9053985.
[34] F. Medhat, D. Chesmore, J. Robinson, Masked conditional neural networks for automatic sound events recognition, in: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2017, pp. 389–394. doi:10.1109/DSAA.2017.43.
[35] J. Serrano-Guerrero, J.A. Olivas, F.P. Romero, E. Herrera-Viedma, Sentiment analysis: A review and comparative analysis of web services, Inf. Sci. 311 (2015) 18–38, https://doi.org/10.1016/j.ins.2015.03.040.
[36] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, B. Liu, Combining lexicon-based and learning-based methods for twitter sentiment analysis, HP Laboratories, Technical Report HPL-2011 89.
[37] S. Sazzed, S. Jayarathna, Ssentia: A self-supervised sentiment analyzer for classification from unlabeled data, Mach. Learn. Appl. 4 (2021) 100026, https://doi.org/10.1016/j.mlwa.2021.100026.
[38] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The pushshift reddit dataset, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 14, 2020, pp. 830–839.
[39] U.C. Bureau, 2019: ACS 1-Year Estimates Subject Tables, https://data.census.gov/ (accessed October 5, 2021).
[40] U.C. Bureau, Preventing Cancer Across a Lifetime, https://www.cdc.gov/cancer/dcpc/prevention/lifetime.htm (accessed October 5, 2021).
[41] S.G. Reed, N. Grijebovskaya Duffy, K.C. Walters, T.A. Day, Oral cancer knowledge and experience: a survey of south carolina medical students in 2002, Journal of cancer education 20 (3) (2005) 136–142, https://doi.org/10.1207/s15430154jce2003_6.
[42] E.J. Coups, S.L. Manne, C.J. Heckman, Multiple skin cancer risk behaviors in the us population, American journal of preventive medicine 34 (2) (2008) 87–93, https://doi.org/10.1016/j.amepre.2007.09.032.
[43] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, arXiv preprint arXiv:1908.09635. doi:10.1145/3457607.