Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Exploring Convolutional Neural Networks and Topic Models for User Profiling from Drug Reviews

Elena Tutubalina · Sergey Nikolenko

Received: date / Accepted: date

Abstract Pharmacovigilance, and generally applications of natural language processing models to healthcare, have attracted growing attention over the recent years. In particular, drug reactions can be extracted from user reviews posted on the Web, and automated processing of this information represents a novel and exciting approach to personalized medicine and wide-scale drug tests. In medical applications, demographic information regarding the authors of these reviews such as age and gender is of primary importance; however, existing studies usually either assume that this information is available or overlook the issue entirely. In this work, we propose and compare several approaches to automated mining of demographic information from user-generated texts. We compare modern natural language processing techniques, including extensions of topic models and convolutional neural networks (CNN). We apply single-task and multi-task learning approaches to this problem. Based on a real-world dataset mined from a health-related web site, we conclude that while CNNs perform best in terms of predicting demographic information by jointly learning different user attributes, topic models provide additional information and reflect gender-specific and age-specific symptom profiles that may be of interest for a researcher.

Keywords text mining · natural language processing · topic modeling · deep learning · convolutional neural networks · multi-task learning · single-task learning · user reviews · demographic prediction · demographic attributes · social media · mental health

Elena Tutubalina
Kazan (Volga Region) Federal University, Kazan, Russia
E-mail: elvtutubalina@kpfu.ru

Sergey Nikolenko
Steklov Institute of Mathematics at St. Petersburg, Russia

Kazan (Volga Region) Federal University, Kazan, Russia
E-mail: sergey@logic.pdmi.ras.ru

1 Introduction

The rapidly growing field of *pharmacovigilance* is concerned with healthcare-relevant information that can be collected automatically from publicly available sources. In particular, patients widely publish messages associated with health information online in social media, discussion groups, and message boards. These user-generated texts contain a vast amount of unstructured information that can be directly or indirectly linked to health conditions. This provides latent evidence about the patients' personal health and user attitude, and it has the potential to be a valuable external source of information for health applications ranging from understanding opinions about mental health to detection of adverse drug reactions. However, given that the key objective of pharmacovigilance research is to automatically monitor traditional and modern channels to identify potentially adverse drug events and interactions, it is important to explore the relationship between social media and health-related factors including user characteristics or medical conditions [37].

Important user attributes such as age and gender that are directly useful for providing personalized healthcare services (e.g., signal detection or recommendation) are not typically available. When such information is provided, e.g., when the texts are collected from electronic health records about patients with explicitly known age and gender, there is no problem. However, in many situations user reviews for drugs and medical services are found anonymously on review web sites such as *WebMD.com*; often demographic information is available for a minority of users but not all of them. Hence, the problem arises to predict user demography based on the texts of user reviews.

In this work, we make the first steps in the direction of extracting demographic information from user-generated texts related to medical subjects. We have collected a database of medical reviews from a health-related web site with user-generated content, namely *WebMD*, and have trained models to predict the age and gender of users who wrote these reviews. We compare three types of models: topic models with user attributes such as PLDA and USTM, and neural models based on CNNs on top of *word2vec* embeddings. This work is a significantly extended journal version of the conference paper [67]; compared to the conference version, we have significantly extended our experimental part with topic models and convolutional neural networks based on single-task and multi-task architectures. Note also another extension of the same conference paper [68], where we developed a feature-rich machine learning approach with new domain-specific information aiding the classifiers.

The paper is organized as follows. In Section 2, we survey related work about demographic prediction and natural language processing in the biomedical or health domain. In Section 3, we present our models for both unsupervised and supervised learning that we compare in this work, namely partially labeled topic models and a convolutional architecture for text classification. We present experimental results in Section 4 and conclude with Section 5.

2 Background

In this section, we review existing studies on demographic prediction and applications of Natural Language Processing (NLP) in biomedical domain and healthcare.

2.1 Demographic Prediction

The effects of gender, age, social class, and other individual attributes on conversational discourse have been widely investigated in sociolinguistic research [17, 24]. Most research has focused on modeling the diversity of speaking and writing styles associated with various demographic attributes. Starting with the pioneering work of Rao et al. [54] on author-property discovery in microblogs, computational models have been focusing on the classification of latent user attributes in text generated on social media platforms such as *Twitter* or *Facebook*. Several studies have discovered correlations between expressions of subjectivity and gender and leveraged these correlations for gender identification [12, 54, 58]. Recent approaches have incorporated gender differences to improve sentiment classification in social media [71, 74]. In particular, Yang et al. [74] proposed *User-Aware Sentiment Topic Models* (USTM) which incorporate user metadata (e.g., gender, age, or location) with topics and sentiments in an unsupervised manner. In this model, topics depend on the document's tags, and words are conditioned on the latent topics, sentiments and tags. USTM gave a substantial improvement over other sentiment topic models which do not incorporate additional demographic information in the task of predicting review sentiment.

Most prior research on age prediction has defined this task as a two-class or multi-class classification problem [27, 54, 67, 68]. In [27], Garera and Yarowsky evaluated several methods for the classification of gender, age, and native language in conversations and emails. They explored a variety of novel sociolinguistic and discourse-based features. Experiments showed significant performance gains from the joint modeling of speaker attributes along with partner attributes. In [44], Nguyen et al. presented a linear regression approach training on three corpora: blogs, transcribed telephone speech, and posts from an online forum on breast cancer. They conducted that differences between the corpora are reflected in a set of effective features. The effective features are about being pregnant and having kids and about story telling nature of posts in the cancer and blog datasets, respectively. Alekseev and Nikolenko studied the problem of age prediction based on user-generated texts on a Russian-language dataset [2, 3], proposing several algorithms that operate on word embeddings on social network statuses of the users. In [54], Rao et al. showed that the use of emoticons and sociolinguistics-inspired features did not result in an improvement in the age prediction in tweets, unlike gender prediction where it was actually useful.

2.2 Text Mining in the Biomedical Domain and Healthcare

One development of recent years is the rise of the Internet of Things (IoT) [6], with healthcare among its most prominent applications. While IoT solutions for healthcare usually deal with special sensors that collect health-related information directly [11,33,48], there is certainly a place for text mining in IoT solutions as well. Chat bots can actively talk to patients in order to collect relevant information, but an even better idea might be to use passive solutions that monitor the texts on social media for health-related information; such studies also often come under the label of pharmacovigilance.

In [37], Kotov surveyed recent work on social media analytics for healthcare that demonstrates the capability of social media data for mining health-related knowledge. In [13], the use of *Facebook* was investigated for gathering medical data from young subjects while promoting healthier habits. Automatic event extraction form text is an important step in knowledge acquisition [64]. Therefore, the majority of biomedical and clinical NLP research on social media use has solely focused on identification of entities such as adverse drug events and health conditions, entity-entity relations, medical concept normalization [4,15,41,42,70] and understanding opinions through sentiment analysis [14,16,21,22,30,43,49,59,69,73].

Sentiment analysis is used to solve various problems in the field of medicine. For example, in [57] sentiment analysis was used to evaluate the effectiveness of treatment. These studies are necessary for doctors to make a decision about patient's treatment. Participation of patients in making decisions about the treatment and management of their own medical care can lead to an improvement the quality of the provided medical services and the results of treatment [20]. Cambria et al. presented the Sentic PROMs system, which assessed the quality of healthcare [14]. The system collects user feedback and extracts emotions in them. This allows monitoring the quality of care and treatment in medical institutions in a real time. In [72], a multi-step approach to the classification of patient opinions from the Patient Opinion service[1] was proposed. In [21,22], classical sentiment analysis was applied to classify of medical records into negative and positive classes in order to determine the effectiveness of patient treatment. Often, the sentiment analysis applies to reviews about medical goods and drugs. The works [43,59,73] analyzed the sentiment of patient reviews about various drugs, and the work [1] evaluated different medical devices for the treatment of joint and muscle pain. These studies are conducted, first, in order to identify side effects, and second, as a tool for the pharmaceutical companies producing the drugs to find out advantages and defects of their products, improve and promote their drugs. Sentiment analysis was also used to evaluate the psychological state of patients: Salas-Zrate et.al. evaluated the sentiment of the tweets associated with diabetes [56], and in [9, 47, 55] machine learning was used to classify records from the social network of people with cancer in order to identify their general mood.

[1] http://www.patientopinion.org.uk

Other models deal with the temporal evolution of tweets or other user-generated texts. In [40], a basic ARIMA model is used to monitor public health discussions, while an interesting recent work [60] uses temporal extensions of topic models, specifically TemporalLDA and the proposed Temporal Ailment Topic Aspect model intended to monitor public health on *Twitter* over time. In [39], Liu et al. investigated semantic representations based on Latent Dirichlet Allocation (LDA) for categorization and representation problems in community-based health services.

Several groups started in the last years with considering language-derived personality and demographic information to analyze diagnoses of mental illness [30, 49]. As recently reviewed in [18], applications of social media-based NLP to the mental health domain can address public health research questions. Experimental results on various corpora (e.g., a dataset of depression-indicative tweets) showed fair to good correlations between personality scores, linguistic features, and official statistics provided by US centers for disease controls. In [41], manual analysis showed that women that describe adverse drug reactions such as abnormal pain are more emotional than patients for whom muscle pain and spasms are symptoms of the underlying disease. In [49], Preotiuc-Pietro et al. analyzed tweets from users who share their mental illness such as depression and post traumatic stress disorder (PTSD). Their dataset contained 370 and 483 users diagnosed only with PTSD and depression, respectively. The experiments showed that gender is weakly predictive of any mental illness, age is highly predictive for PTSD users, and both age and gender contain complementary information. In [8], Benton et al. examined multi-task learning techniques on mental health tasks. Their primary objective was to predict a health condition. Conditions are classified into the following classes: neurotypicality, anxiety, depression, suicide attempt, eating disorder, panic attacks, schizophrenia, bipolar disorder, and PTSD. An auxiliary task was gender prediction. Experimental results of multi-task feedforward networks with two hidden layers showed that multi-task networks trained on both main and auxiliary tasks achieved more accurate results on condition prediction and lower results on predicting gender itself than single-task networks. In [45], Nguyen et al. explored the textual cues of online communities interested in depression extracting psycholinguistic features and content topics.

Chou et al. analyzed sociodemographic and health-related factors associated with social media users in the United States [16]. Results of this study support the conclusion that health communication programs must consider the age of the particular population and forms of social media (e.g., blogs, social networking sites), while ethnic and health statusrelated user characteristics do not affect social media use. This leads to the need to mine demographic information about the authors together with the user-generated texts themselves.

In our previous works [67, 68, 70], we provided first results on automated predictions of demographic attributes based specifically on medical reviews. In [67], our experimental results showed that convolutional architectures work better than recurrent networks of LSTMs for both age and gender prediction. In [68], we experimented with baseline machine learning methods as well as

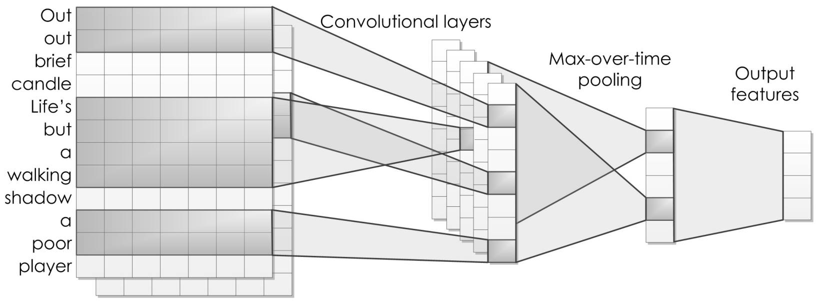

Fig. 1: A convolutional neural network with 1D convolutions over a text.

feature-rich machine classifiers, and in [70], combined deep recurrent neural networks with conditional random fields to extract adverse drug reactions from free-form text. Although these studies shed some light on how to predict demographic information, joint prediction of age and gender has not been evaluated through multi-task learning. Therefore, this study can be viewed as an extension of the previous work.

3 Models

In this section, we introduce the models we use for our user profiling approaches. They can be divided into two major groups of models:

(1) Convolutional neural networks with single-task and multi-task learning architectures;
(2) User-aware topic models that can learn intermediate distributions in an unsupervised manner.

Sections 3.1 and 3.2 describe these models respectively.

3.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) consider inputs where there is some notion of "spatial distance" between input dimensions. In particular, CNNs were designed primarily for computer vision problems, where the input image has a natural notion of distance between the pixels, and it is natural that nearby pixels in an image have a much stronger relation to each other and are much more likely to belong to the same object than distant pixels. In a typical CNN, each neuron on the next layer is connected not with all, but only with a small localized subset of neurons on the previous layer; such *convolution* layers usually alternate with *pooling* layers, where activations from different neurons are pooled together. This approach resembles how the human visual cortex actually works.

The basic ideas are illustrated on Fig. 1 for the case of a one-dimensional CNN, characteristic for natural language processing. Layers are connected in a sparse way: units on level k receive as input only a subset of units on level $k-1$. At the same time, each filter on a hidden layer is replicated across the entire input vector, learning the same localized features in every part of the input; this means that the weights are shared, and the total number of parameters is not so overwhelming. A *feature map* thus represents repeated applications of the same unit across all local neighborhoods, i.e., a convolution of the input with a linear filter followed by a nonlinearity; a single hidden layer can contain several feature maps. Convolutional layers are usually interleaved with *pooling*, or *subsampling* layers that combine subsets of the input and output the maximum values of all features; here the idea is that a higher-level feature's exact location is less important than its interaction with other neighboring features; in one-dimensional CNNs, these are usually *max-over-time* pooling layers, which output the maximal value of a feature map along a window.

CNNs are a natural fit for image processing, commonly applied to such problems as image classification and recognition, character recognition, image segmentation and object recognition, scene labeling, video processing, and so on [38, 53]. As an example of applying CNNs to NLP, we consider a rather vanilla application of CNNs for semantic sentence classification [35]. The model is not as deep as computer vision models, with only one convolutional layer with max-over-time pooling and a softmax output; regularization is achieved through dropout, with consistent and significant improvement in accuracy with dropout across all experiments; the model is trained on prepared *word2vec* word embeddings and does not attempt to tune word representations for better results. Still, the authors report better results on such tasks as sentiment analysis and sentence classification than baseline techniques that include recursive autoencoders and recursive neural networks with parse trees.

3.1.1 Single-task Learning and Multi-task Learning

We experiment with a single-task learning (STL) architecture and a multi-task learning (MTL) architecture based on convolutional neural networks. Recent studies [8, 26, 63] have indicated that predicting several related tasks should allow the model to better exploit correlations between the predictions.

The first configuration shown in Figure 2a is the STL model. The input layer is based on pre-trained word embeddings. This layer is fed into CNN similar to the one recently presented in [35]. CNN's output is fed into a fully connected layer with dense connections and hyperbolic tangent as the activation function. Finally, the dense layer's output is fed to into a fully connected layer with softmax activation. The second configuration, shown in Figure 2b, is the MTL model, where the hidden layers of CNN are shared between the two tasks, i.e., age and gender predictions. An additional per-task hidden dense layer is used to give the model flexibility to map from the shared representation to a task-specific one.

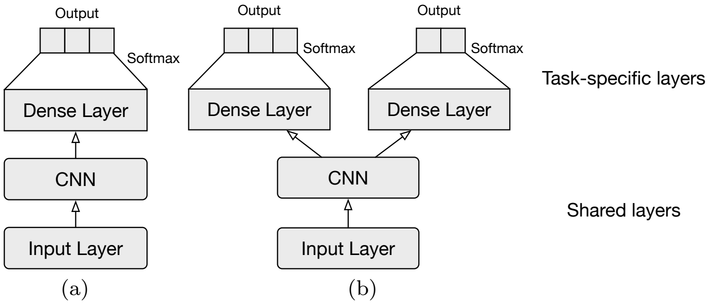

Fig. 2: Two architectures: (a) Singe-task learning; (b) Multi-task learning.

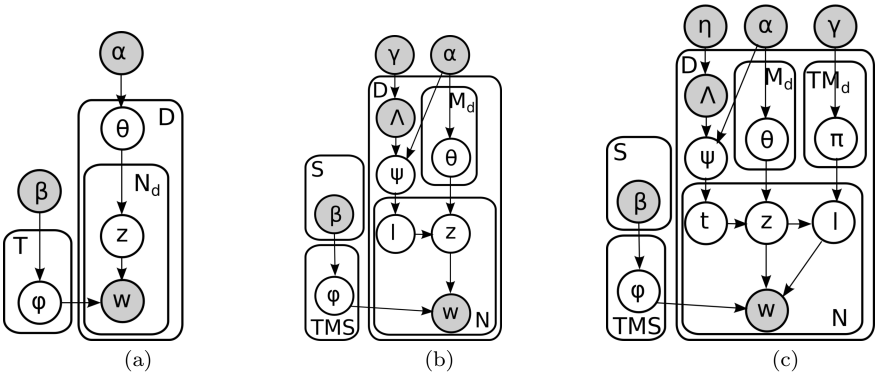

Fig. 3: (a) LDA; (b) PLDA; (c) USTM.

3.2 Topic Modeling

We assume that a corpus of D documents contains T topics expressed by W different words. Each document $d \in D$ is modeled as a discrete distribution $\theta^{(d)}$ on the set of topics: $p(z_w = t) = \theta_{td}$, where z is a discrete variable that defines the topic of each word $w \in d$. Each topic, in turn, corresponds to a multinomial distribution on words: $p(w \mid z_j = t) = \phi_{wt}$ (here w denotes words in the vocabulary and j denotes individual instances of these words). The model introduces Dirichlet priors with parameters α for topic vectors θ, $\theta \sim \mathrm{Dir}(\alpha)$, and β for word distributions ϕ, $\phi \sim \mathrm{Dir}(\beta)$.

To process medical reviews, we employ two extensions of latent Dirichlet allocation (LDA) model:

– Partially Labeled Topic Model (PLDA) [51];
– User-aware Sentiment Topic Models (USTM) [74].

The graphical model of LDA [10, 29] is shown on Figure 3a. In the basic LDA model, a document is generated word by word: for each word, first sample its topic index t from θ_d, $t \sim \mathrm{Mult}(\theta_d)$, then sample the word w from ϕ_t, $w \sim \mathrm{Mult}(\phi_t)$. We denote by $n_{w,t,d}$ the number of words w generated with topic t in document d; partial sums over such variables are denoted by asterisks, e.g., $n_{*,t,d} = \sum_w n_{w,t,d}$ is the number of all words generated with topic t in document d, $n_{w,*,*} = \sum_{t,d} n_{w,t,d}$ is the total number of times word w occurs in the corpus and so on; we denote by $\neg j$ a partial sum over "all instances except j", e.g., $n^{\neg j}_{w,t,d}$ is the number of times word w was generated by topic t in document d except position j (which may or may not contain w). Inference proceeds with *collapsed Gibbs sampling*, where θ and ϕ variables are integrated out, and z_j are iteratively resampled as follows:

$$p(z_j = t \mid \boldsymbol{z}_{-j}, \boldsymbol{w}, \alpha, \beta) \propto \frac{n^{\neg j}_{*,t,d} + \alpha}{n^{\neg j}_{*,*,d} + T\alpha} \cdot \frac{n^{\neg j}_{w,t,*} + \beta}{n^{\neg j}_{*,t,*} + W\beta},$$

where $\boldsymbol{z}_{-j}$ denotes the set of all z values except z_j.

3.2.1 Partially Labeled and User-aware Sentiment Topic Models

A further extension of the PLDA model is presented in the recently developed User-aware Sentiment Topic Models (USTM) [74]; USTM incorporates user meta-data tags (e.g., location, gender, or age) together with topics and sentiment. In this model, each document is assigned with an observed tag or a combinations of tags, topics are generated conditioned on the document's tags, sentiment labels are generated conditioned on the (document, tag, topic) triples, and words are conditioned on the latent topics, sentiments and tags. Formally, a tag distribution ψ_d is generated for every document (with a Dirichlet prior η), for each position j a tag $a_j \sim \mathrm{Mult}(\psi_d)$ is drawn from ψ_d, and distributions of topics, sentiments, and words are conditional on the tag a_j. The USTM graphical model is shown on Fig. 3(c). Denoting by $n_{w,k,t,m,d}$ the number of words w generated with topic t, sentiment label k, and metadata tag m in document d and extending the notation accordingly, a Gibbs sampling step proceeds as

$$p(z_j = t, l_j = k, a_j = m \mid \nu) \propto$$
$$\frac{n^{\neg j}_{*,*,t,m,d} + \alpha}{n^{\neg j}_{*,*,*,*,d} + TM_d\alpha} \cdot \frac{n^{\neg j}_{w,*,t,m,*} + \beta}{n^{\neg j}_{*,*,t,m,*} + W\beta} \times \frac{n^{\neg j}_{w,k,t,m,*} + \beta_{wk}}{n^{\neg j}_{*,k,t,m,*} + \sum_w \beta_{wk}} \cdot \frac{n^{\neg j}_{*,k,t,m,d} + \gamma}{n^{\neg j}_{*,*,t,m,d} + S\gamma},$$

where M_d is the number of tags in document d.

Samples are then used to estimate model variables:

$$\theta_{dmt} = \frac{n_{*,*,t,m,d} + \alpha}{n_{*,*,*,m,d} + T * \alpha} \tag{1}$$

$$\pi_{dmtk} = \frac{n_{*,k,t,m,d} + \gamma}{n_{*,*,t,m,d} + S * \gamma} \tag{2}$$

$$\phi_{jksw} = \frac{n_{w,k,t,m,*} + \beta^{wk}}{\sum_{w=1}^{V} n_{w,k,t,m,*} + \sum_w \beta_{wk}} \tag{3}$$

For each review d, the distribution over user attributes $P(a|d)$ can be estimated using sentiment-based topics for each tag group, ϕ, by marginalizing out the topics t, and sentiments, l, as follows:

$$P(a|d) \propto P(d|a) = \prod_{w \in \boldsymbol{w}_d} P(w|a) = \prod_{w \in \boldsymbol{w}_d} \sum_{t=1}^{T_a} \sum_{l=1}^{S} P(w|l,t,a) \tag{4}$$

For comparison, we also use a predecessor of USTM that can be though of as a simplified version of USTM, namely *Partially Labeled Topic Model* (PLDA) [51]. PLDA operates exactly the same as USTM, but without a separate latent variable for sentiment labels l_j. Its graphical model is shown on Fig. 3b, and the Gibbs sampling step proceeds as

$$p(z_j = t, a_j = m \mid \nu) \propto \frac{n_{*,t,m,d}^{\neg j} + \alpha}{n_{*,*,*,d}^{\neg j} + TM_d\alpha} \cdot \frac{n_{w,t,m,*}^{\neg j} + \beta}{n_{*,t,m,*}^{\neg j} + W\beta} \cdot \frac{n_{w,t,m,*}^{\neg j} + \beta_{wk}}{n_{*,t,m,*}^{\neg j} + \sum_w \beta_{wk}}$$

For PLDA, the distribution over user attributes $P(a|d)$ can be estimated using ϕ by marginalizing out the topics t as follows:

$$P(a|d) \propto P(d|a) = \prod_{w \in \boldsymbol{w}_d} P(w|a) = \prod_{w \in \boldsymbol{w}_d} \sum_{t=1}^{T_a} P(w|t,a) \tag{5}$$

For all models, posterior inference was done with 1000 Gibbs iterations with hyperparameters $\alpha = 50/K$, $\gamma = 0.01$, K is the total number of topics. For USTM, we adopted the publicly available sentiment lexicon MPQA [65] which is often used in real world sentiment analysis problems as the seed dataset of sentiment words. We divided sentiment priors into three different values (neutral, positive, and negative) and set the β priors for all words in the corpus similarly to [66, 74].

4 Evaluation and Experiments

In this section, we present our experiments with the described models.

4.1 Dataset

Experimental evaluation of the proposed models was conducted on a real world data set. This data set (further referred to as **WebMD**) consists of reviews crawled from WebMD.com[2], a health information services website that aims to provide credible and trustworthy information, supportive communities, and in-depth material about health subjects. We have crawled 217,485 reviews from

[2] `http://www.webmd.com`

Table 1: Summary statistics for the WebMD dataset. The number in parentheses indicates the number of reviews associated with a label.

Top-5 Conditions	**Gender**	**Age groups**
high blood pressure (10201)	Female (23343)	45-54 (8430)
pain (9306)	Male (9979)	55-64 (7056)
depression (7340)		35-44 (6207)
chronic trouble sleeping (3454)		19-34 (7410)
attention deficit disorder with hyperactivity (3021)		65 or over (4219)

authors tagged as “Patient” on *WebMD*. Each review contains the following fields:

(1) Date when the review was written;
(2) Condition for taking treatment;
(3) Textual review given for the effects caused by the use of the drug;
(4) User attributes.

Review authors select conditions from a predefined set for each drug. In order to avoid the sparsity issue and and exclude conditions with bias (e.g., “Pregnancy”, “Premenstrual Disorder with a State of Unhappiness”), we selected reviews associated with 5 most commented conditions for training/testing.

Each review is associated with a gender tag (“Male” or “Female”) and one of predefined age tags: “19–24”, “25–34”, “35-44”, “45-54”, “55-64”, “65-74”, or “75 or over”. To examine differences by age in the relationship between personal health-related age factors and social media texts, we divided all authors’ age tags into three groups: the young adults (ages 18-34), the middle-age group (ages 35-64), and the aged (65 and older). This separation chunks is based on psychological and medical science research that defines young adults as those ages between 18 and 29 [5, 19, 25, 32, 61]. Meanwhile, the medical treatment for those patients 65 years of age and older differ considerably from those for the younger patients [28, 62].

Table 1 presents summary statistics for the WebMD dataset. In these 33,332 documents, the total vocabulary size was 40,728, and we found between 20% to 30% of the words from this vocabulary in the standard pretrained *word2vec* models (the vectors are described below); most of the other words were typos and misspellings, and we did not try to correct for them. We performed pre-processing by lower-casing all words.

4.2 Neural Networks’ Settings

In order to get local features from a review with CNNs we have used multiple filters of different lengths [35]. We used a sliding max-pooling window of length 2 to get features through filters. Pooled features are then fed to a fully connected feed-forward neural network (with dimension 100) to make inference using hyperbolic tangent (tanh) as activation function. Then we apply a layer with the softmax activation with number of outputs equal to number

Table 2: Statistics of *word2vec* embeddings

Embeddings	Dimension	# of tokens	# of tokens in WebMD vocab.
NewsVec	300	3,000,000	29,161
TwitterVec	200	1,859,182	33,891
WikiVec	200	955,839	27,730
PubmedWikiVec	200	5,443,656	314,79

of classes. We applied the Adam optimization algorithm [36] and used the dropout rate of 0.5 after the embedding layer (before convolutional layers). Embedding layers are trainable for all networks. We set the mini-batch size to 128. For out-of-vocabulary words with the pre-trained word model, we used representations uniformly sampled from the range of embedding weights [31].

In order to choose the best number of epochs for training, we separated out 10% of the training dataset to form the validation set. We employed early stopping after two epochs with no improvement on the validation set. After there experiments, we set the number of the training epochs to 8 for all models.

We tested the following word embeddings using *word2vec* models:

(1) **NewsVec**: word vectors *GoogleNews-vectors-negative300*[3] trained on part of Google News dataset (about 100 billion words);
(2) **WikiVec**: word vectors trained on English Wikipedia [23];
(3) **TwitterVec**: word vectors trained on 200 million English tweets [23];
(4) **PubmedWikiVec**: word vectors trained on biomedical scientific literature from PubMed, PubMed Central (PMC), and Wikipedia [50].

Basic statistics for the word embeddings are presented in Table 2.

4.3 Experiments and Results

We evaluate our model by comparing with several machine learning methods:

- Baseline classifier with the following features: occurrence of contiguous sequences of 1-, 2-, and 3-grams; the maximum number of features are 25,000. The machine learning approaches are Support Vector Machines (SVM) and Logistic regression.
- Feature-rich classifier proposed which leverages a variety of baseline, surface-form, semantic, cluster-based, distributed and lexicon features described in our previous work [68]. The lexicon features are based on (i) sentiment lexicons of positive and negative words and (ii) dictionaries of adverse drug reactions.

For evaluation, we performed 5-fold cross-validation on the selected data. We evaluated the performance of our models using standard text classification metrics: precision (P), recall (R), F_1-measure (F1).

The macro-averaged results are shown in Tables 3 and 4. As follows from both tables, the proposed neural models with MTL architecture can be used

[3] https://code.google.com/archive/p/word2vec/

Table 3: Gender prediction (macro-averaged, 2 classes)

Method	**P**	**R**	**F1**
Unsupervised models			
PLDA (20 topics)	0.634	0.654	0.636
PLDA (70 topics)	0.639	0.656	0.644
PLDA (100 topics)	0.647	0.654	0.650
USTM (20 topics)	0.646	0.621	0.628
Single-task models (trained only for gender prediction)			
SVM with baseline features	0.645	0.649	0.647
Feature-rich SVM	0.674	0.676	0.675
Logistic regression with baseline features	0.671	0.662	0.666
Feature-rich Logistic regression	0.702	0.691	0.695
CNN, PubmedWikiVec, [2, 3]	0.692	0.667	0.674
CNN, TwitterVec, [2, 3]	0.690	0.668	0.674
Multi-task models (joint models for both age and gender prediction)			
CNN, PubmedWikiVec, [2, 3]	0.696	0.676	0.683
CNN, TwitterVec, [2, 3]	0.689	0.683	0.684
CNN, WikiVec, [2, 3]	0.690	0.669	0.675
CNN, NewsVec, [2, 3]	0.691	0.674	0.680

to predict the attributes of review authors with reasonable accuracy. Furthermore, analysis of the presented results leads to three important conclusions.

First, the proposed models with MTL architecture consistently perform better on gender prediction, while staying roughly on par on age prediction. Second, multi-task CNNs outperform single-task CNNs in terms of recall. The multi-task CNNs trained on PubmedWikiVec outperform the feature-rich SVM on gender prediction task and both machine learning classifiers on age prediction. the best results on gender prediction have been obtained while using vectors trained on social media posts or on texts from PubMed, PMC and Wikipedia. The best results on age prediction have been obtained while using vectors trained on a corpus of news. Third, PLDA shows better performance in predicting user attributes than USTM. This result is, in our opinion, due to the dataset size which in this case is probably not large enough for USTM with the larger number of learned distributions to shine.

Table 5 provides an additional in-depth comparison between several different sets of filters for the CNNs. Several important conclusions can be derived based on the results in Table 5. First, the models without dropout between the CNN layers and the dense layer significantly outperform the models with dropout on both tasks. Second, the models based on layers with hyperbolic tangent (tanh) activation functions significantly outperform models with rectified linear unit (ReLU) as activation functions. Third, tuning the set of CNN filters helps to improve the networks' classification performance on both tasks, but it is clear that all sufficiently expressive architectures work well enough.

The results of CNN on texts associated with different health conditions shown in Tables 6 and 7. We can make the following observations. First, CNN achieved the highest results in gender classification on texts about ADHD. As concluded in [52], there are clear gender differences with respect to the

Table 4: Age prediction (macro-averaged, 3 classes)

Method	P	R	F1
Unsupervised models			
PLDA (20 topics)	0.468	0.543	0.463
PLDA (70 topics)	0.484	0.539	0.490
PLDA (100 topics)	0.496	0.528	0.500
USTM (20 topics)	0.523	0.357	0.321
Single-task models (trained only for age prediction)			
SVM with baseline features	0.514	0.513	0.513
Feature-rich SVM	0.540	0.539	0.539
Logistic regression with baseline features	0.562	0.521	0.536
Feature-rich Logistic regression	0.574	0.544	0.557
CNN, PubmedWikiVec, $[2, 3]$	0.593	0.548	0.561
CNN, TwitterVec, $[2, 3]$	0.592	0.554	0.566
Multi-task models (joint models for both age and gender prediction)			
CNN, PubmedWikiVec, $[2, 3]$	0.584	0.563	0.566
CNN, TwitterVec, $[2, 3]$	0.588	0.555	0.566
CNN, WikiVec, $[2, 3]$	0.578	0.556	0.561
CNN, NewsVec, $[2, 3]$	0.573	0.578	0.574

Table 5: Evaluation of different CNN parameters, PubmedWikiVec, the hyperbolic tangent (tanh) activation function, multi-task learning.

CNN	Gender prediction			Age prediction		
	P	R	F1	P	R	F1
$[1, 2]$, tanh, w/o dropout after CNN	0.690	0.669	0.677	0.569	0.554	0.557
$[2, 4]$, tanh, w/o dropout after CNN	0.706	0.665	0.676	0.586	0.554	0.566
$[2, 3]$, tanh, w/o dropout after CNN	0.696	0.676	0.683	0.593	0.548	0.561
$[2, 3]$, ReLU, w/o dropout after CNN	0.701	0.663	0.674	0.596	0.548	0.564
$[1, 2, 3]$, tanh, w/o dropout after CNN	0.691	0.673	0.679	0.575	0.554	0.559
$[2, 3, 4]$, tanh, w/o dropout after CNN	0.698	0.671	0.675	0.591	0.557	0.568
$[1, 2, 3, 4, 5]$, tanh, w/o dropout after CNN	0.698	0.664	0.674	0.593	0.561	0.573
$[2, 3]$, tanh, w/ dropout after CNN	0.687	0.651	0.661	0.584	0.514	0.534
$[2, 3]$, ReLU, w/ dropout after CNN	0.703	0.632	0.642	0.591	0.468	0.491
$[2, 3, 4]$, tanh, w dropout after CNN	0.695	0.655	0.665	0.565	0.540	0.551
$[1, 2, 3, 4]$, tanh, w/ dropout after CNN	0.685	0.659	0.667	0.578	0.533	0.545
$[2, 3, 4, 5]$, tanh, w/ dropout after CNN	0.686	0.662	0.671	0.580	0.538	0.551
$[1, 2, 3, 4, 5]$, tanh, w/ dropout after CNN	0.690	0.662	0.670	0.580	0.544	0.553

prevalence of ADHD. Much lower results are achieved on texts about another mental health disorder, i.e. depression. As shown in [34], it is difficult to define gender-specific symptom profiles in men and women even if there has been increased awareness of gender-specific issues in depression in recent years. Second, CNN achieved the highest results in age prediction on texts about chronic trouble sleeping. Since sleep problems affect all age groups, there are several groups of age-specific patterns linked with educational level, perceived health, mood, smoking habits, etc. [7].

In order to analyze the difference between results associated with a particular disorder, Tables 8 and 9 present topic examples derived from texts on

Table 6: Gender prediction on texts associated with different health conditions; evaluation of CNN (multi-task learning, PubmedWikiVec, [2, 3]).

Health condition	**P**	**R**	**F1**
Pain	0.682	0.669	0.673
Depression	0.676	0.651	0.661
High Blood Pressure	0.699	0.682	0.687
Chronic Trouble Sleeping	0.671	0.650	0.656
Attention Deficit Disorder with Hyperactivity (ADHD)	0.733	0.706	0.715

Table 7: Age prediction on texts associated with different health conditions; evaluation of CNN (multi-task learning, PubmedWikiVec, [2, 3])

Health condition	**P**	**R**	**F1**
Pain	0.560	0.520	0.533
Depression	0.553	0.504	0.516
High Blood Pressure	0.528	0.487	0.500
Chronic Trouble Sleeping	0.565	0.529	0.540
Attention Deficit Disorder with Hyperactivity (ADHD)	0.528	0.532	0.528

Table 8: Sample topics discovered by PLDA for the gender tags.

#	**Topic words about attention deficit disorder with hyperactivity**
	male
1	life med drugs son young family early behavior research changing addictive wife childhood produce box problem reason quality anti warnings
2	pain stomach started med male skin patch food empty heartburn minutes basis expected fair abdominal eaten quickly upset penis takes absolutely making
3	side dexedrine amphetamines adderall pretty physical evekeo effects feel told benefits interest treatment kid weird dexdro years ill makes past lose fact
4	adderal worse effective generic effects crash condition insurance top works cost bad face conditions evekeo effect pharmacy tough spine mention
	female
1	work school started son diagnosed family mom meds kids med husband house home friends give behavior things love adhd person working child changed
2	skin hands face early rash problems picking feet fingers legs doctors wanted fast adderral patch eat strange cold art accomplish issues suffered nails arms
3	feel med makes headaches tired asleep fall stomach head upset hyper started lack fast long jittery thinking sit calm world give aches working hope faster
4	taking pain adderal hair worse stopped thing head normal neck back horrible loss weird lexapro makes lot today fact changing sort difference falling stop

reviews about a particular condition using PLDA based on a unigram representation of reviews. For each topic, we report terms with highest weights. Quantitative evaluation of topic models is a known open problem [46], but we can make several qualitative observations based on the results in both tables. First, we observe differences in expressions about how ADHD typically affects women and men. Men with ADHD tend to have more problems with stomach and addiction issues compared to women with ADHD. Women experience skin problems, headaches, nail biting. Second, both males and females are talk-

Table 9: Sample topics discovered by PLDA for the age tags.

#	Topic words about chronic trouble sleeping
	the young adults (ages 19-34)
1	work made morning dream long safe wonders sleepy end addiction drug aid hard knock things downside putting sleep history life lucky fear scary realized
2	medication time ill immediately easy couldnt case recommended fun knocked restless dog easier schedule happened slow physically satisfied treat
3	bed minutes hallucinations laying matter room ready turned crazy happen finally watching needless freaked eyes window starting wasnt benzos hear
	the middle-age group (ages 35-64)
1	car memory driving hospital drove great home hurt dangerous medication fell caused market events isnt blood light walking knew concussion accident
2	medications doc anxiety drugs taking days highly addictive life prescription ambien care person hard dangerous sick including god nightmare withdrawal
3	ativan dose people months happy low mood menopause zoloft dont back slowly med trazadone stay time shot start walk prescribe perfect couldn addicting
	the aged group (65 and older)
1	legs restless rls itch urinate arms subsequent tartrate syndrome weak decline alprazolam history panic acuity thought crazy morphine point personality
2	sleep aids counter aid losing allergy years addictive miracle tylenol recently dangerous figure mind melatonin legs told agitated daughter habit forming
3	feet muscle cramps weakness experiencing leg baby upper hands causing trazodone tension rash painful nervousness intend trust product heart fine show

ing about family-related situations (“kids”, “husband”, and “family”). Third, we observe that chronic sleep disorders affect the mood (“happy”, “scary”, “lucky”) and energy level (“sleepy”, “hard”, “knock”) in people of all ages. The restless legs syndrome becomes more intense for older people (“legs”, “restless”, “rls”, “syndrome”). Examples for the middle-age group show that sleep deprivation leads to accidents or injuries (“car”, “memory”, “driving”, “drove”, “dangerous”). Finally, in all of these examples we see that topic models convincingly show that social media posts contain variable information related to patient experience, and topic models can help us uncover this information from large text streams.

5 Conclusion

In this work, we have tackled a problem often overlooked in pharmacovigilance studies: can we reliably predict the demographic attributes (in our case, age and gender) of a user from his or her review and associated information? We have presented several machine learning approaches to this problem, including a multi-task convolutional neural architecture based on word embeddings and special modifications of topic models. As a result, we have seen that while neural networks in this kind of NLP-related problems perform better in terms of the classification/regression objective, topic models learn and provide extra information that may lead to interesting observations relevant to the underlying healthcare application. We believe that in the future, the interpretability of topic models will be joined with the predictive power of neural networks, and combinations of these approaches, tailored specifically for problems such as prediction of demographic attributes, represent an important direction for further work.

Acknowledgements

This work was supported by the Russian Science Foundation grant no. 15-11-10019. The authors are grateful to Prof. Valery Solovyev for his continuous support. The authors also thank Ilseyar Alimova for her suggestions on related work.

References

1. Adams, D.Z., Gruss, R., Abrahams, A.S.: Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. International Journal of Medical Informatics **100**, 108–120 (2017)
2. Alekseev, A., Nikolenko, S.I.: Predicting the age of social network users from user-generated texts with word embeddings. In: Artificial Intelligence and Natural Language Conference (AINL), IEEE, pp. 1–11. IEEE (2016)
3. Alekseyev, A., Nikolenko, S.I.: Word embeddings of user profiling in online social networks. Computacin y Sistemas **21**(2), 203–226 (2017)
4. Alimova, I., Tutubalina, E.: Automated detection of adverse drug reactions from social media posts with machine learning. In: Proceedings of International Conference on Analysis of Images, Social Networks and Texts (2017)
5. Arnett, J.J.: Emerging adulthood: A theory of development from the late teens through the twenties. American psychologist **55**(5), 469 (2000)
6. Atzori, L., Iera, A., Morabito, G.: The internet of things: A survey. Computer Networks **54**(15), 2787 – 2805 (2010). DOI http://dx.doi.org/10.1016/j.comnet.2010.05.010. URL //www.sciencedirect.com/science/article/pii/S1389128610001568
7. Bardel, A., Wallander, M.A., Wedel, H., Svärdsudd, K.: Age-specific symptom prevalence in women 35–64 years old: A population-based study. BMC Public Health **9**(1), 37 (2009). DOI 10.1186/1471-2458-9-37. URL https://doi.org/10.1186/1471-2458-9-37
8. Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data. In: Proceedings of the 15th Conference of the EACL, vol. 1, pp. 152–162 (2017)
9. Biyani, P., Caragea, C., Mitra, P., Zhou, C., Yen, J., Greer, G.E., Portier, K.: Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 413–417. ACM (2013)
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3**(4–5), 993–1022 (2003)
11. Bui, N., Zorzi, M.: Health care applications: A solution based on the internet of things. In: Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies, ISABEL '11, pp. 131:1–131:5. ACM, New York, NY, USA (2011). DOI 10.1145/2093698.2093829. URL http://doi.acm.org/10.1145/2093698.2093829
12. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309. Association for Computational Linguistics (2011)

13. Buzzi, M.C., Buzzi, M., Franchi, D., Gazzè, D., Iervasi, G., Marchetti, A., Pingitore, A., Tesconi, M.: Facebook: a new tool for collecting health data? Multimedia Tools and Applications **76**(8), 10,677–10,700 (2017). DOI 10.1007/s11042-015-3190-4. URL `https://doi.org/10.1007/s11042-015-3190-4`
14. Cambria, E., Benson, T., Eckl, C., Hussain, A.: Sentic proms: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. Expert Systems with Applications **39**(12), 10,533–10,543 (2012)
15. Choi, S.P., Lee, S., Jung, H., Song, S.k.: An intensive case study on kernel-based relation extraction. Multimedia Tools and Applications **71**(2) (2014). DOI 10.1007/s11042-013-1380-5. URL `https://doi.org/10.1007/s11042-013-1380-5`
16. Chou, W.Y.S., Hunt, Y.M., Beckjord, E.B., Moser, R.P., Hesse, B.W.: Social media use in the united states: implications for health communication. Journal of medical Internet research **11**(4) (2009)
17. Coates, J.: Women, men and language: A sociolinguistic account of gender differences in language. Routledge (2015)
18. Conway, M., OConnor, D.: Social media, big data, and mental health: current advances and ethical implications. Current opinion in psychology **9**, 77–82 (2016)
19. Correa, T., Hinsley, A.W., De Zuniga, H.G.: Who interacts on the web?: The intersection of users personality and social media use. Computers in Human Behavior **26**(2), 247–253 (2010)
20. Coulter, A., Ellins, J.: The quality enhancing interventions project: patient-focused interventions. London: The Health Foundation (2006)
21. Dang, T.T., Ho, T.B.: Mixture of language models utilization in score-based sentiment classification on clinical narratives. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 255–268. Springer (2016)
22. Deng, Y., Stoehr, M., Denecke, K.: Retrieving attitudes: Sentiment analysis from clinical narratives. In: MedIR@ SIGIR, pp. 12–15 (2014)
23. Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., Jaggi, M.: Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1045–1052. International World Wide Web Conferences Steering Committee (2017)
24. Fischer, J.L.: Social influences on the choice of a linguistic variant. Word **14**(1), 47–56 (1958)
25. Fisher, C.R.: Differences by age groups in health care spending. Health Care Financing Review **1**(4), 65 (1980)
26. Gao Z.and Li, S.H., Zhang, G.T., Zhu, Y.J., Wang, C., Zhang, H.: Evaluation of regularized multi-task leaning algorithms for single/multi-view human action recognition. Multimedia Tools and Applications (2017). DOI 10.1007/s11042-017-4384-8. URL `https://doi.org/10.1007/s11042-017-4384-8`
27. Garera, N., Yarowsky, D.: Modeling latent biographic attributes in conversational genres. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pp. 710–718. Association for Computational Linguistics (2009)
28. Glenn, F.: Surgical management of acute cholecystitis in patients 65 years of age and older. Annals of surgery **193**(1), 56 (1981)
29. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101 (Suppl. 1)**, 5228–5335 (2004)
30. Harman, G.C.M.D.C.: Quantifying mental health signals in twitter. ACL 2014 **51** (2014)
31. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034 (2015)
32. Helmert, U., Merzenich, H., Bammann, K.: The association between educational attainment chronic diseases, and cardiovascular disease risk factors in young adults aged 18 to 29 years: results of the federal health survey 1998. SOZIAL-UND PRAVENTIVMEDIZIN **46**(5), 320–328 (2001)

33. Hossain, M.S., Goebel, S., El Saddik, A.: Guest editorial: advances in multimedia for health. Multimedia Tools and Applications **74**(14), 5205–5208 (2015). DOI 10.1007/s11042-014-2202-0. URL `https://doi.org/10.1007/s11042-014-2202-0`
34. Karger, A.: Geschlechtsspezifische aspekte bei depressiven erkrankungen. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz **57**(9), 1092–1098 (2014). DOI 10.1007/s00103-014-2019-z. URL `https://doi.org/10.1007/s00103-014-2019-z`
35. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014). URL `http://arxiv.org/abs/1412.6980`
37. Kotov, A.: Social media analytics for healthcare pp. 309–340 (2015). URL `http://www.crcnetbase.com/doi/abs/10.1201/b18588-11`
38. LeCun, Y., Kavukcuoglu, K., Farabet, C.: Convolutional networks and applications in vision. In: Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, pp. 253–256. IEEE (2010)
39. Liu, M., Zhang, H., Hu, H., Wei, W.: Topic categorization and representation of health community generated data. Multimedia Tools and Applications **76**(8), 10,541–10,553 (2017)
40. McClellan, C., Ali, M.M., Mutter, R., Kroutil, L., Landwehr, J.: Using social media to monitor mental health discussions – evidence from Twitter. Journal of the American Medical Informatics Association p. ocw133 (2016)
41. Miftahutdinov, Z., Tutubalina, E., Tropsha, A.: Identifying disease-related expressions in reviews using conditional random fields. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii **1**(16), 155–166 (2017)
42. Miftakhutdinov, Z., Tutubalina, E.: Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks. CLEF (2017)
43. Na, J.C., Kyaing, W.Y.M., Khoo, C.S., Foo, S., Chang, Y.K., Theng, Y.L.: Sentiment classification of drug reviews using a rule-based linguistic approach. In: International Conference on Asian Digital Libraries, pp. 189–198. Springer (2012)
44. Nguyen, D., Smith, N.A., Rosé, C.P.: Author age prediction from text using linear regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 115–123. Association for Computational Linguistics (2011)
45. Nguyen, T., O'Dea, B., Larsen, M., Phung, D., Venkatesh, S., Christensen, H.: Using linguistic and topic analysis to classify sub-groups of online depression communities. Multimedia Tools and Applications **76**(8), 10,653–10,676 (2017). DOI 10.1007/s11042-015-3128-x. URL `https://doi.org/10.1007/s11042-015-3128-x`
46. Nikolenko, S.I.: Topic quality metrics based on distributed word representations. In: Proc. 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1029–1032 (2016)
47. Ofek, N., Caragea, C., Rokach, L., Biyani, P., Mitra, P., Yen, J., Portier, K., Greer, G.: Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon. In: Social Intelligence and Technology (SOCIETY), 2013 International Conference on, pp. 109–113. IEEE (2013)
48. Pogorelc, B., Bosnić, Z., Gams, M.: Automatic recognition of gait-related health problems in the elderly using machine learning. Multimedia Tools and Applications **58**(2), 333–354 (2012). DOI 10.1007/s11042-011-0786-1. URL `https://doi.org/10.1007/s11042-011-0786-1`
49. Preotiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H.A., Ungar, L.: The role of personality, age and gender in tweeting about mental illnesses. In: NAACL HLT, vol. 2015, p. 21 (2015)
50. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., Ananiadou, S.: Distributional semantics resources for biomedical text processing. Proceedings of Languages in Biology and Medicine (2013)
51. Ramage, D., Manning, C.D., Dumais, S.: Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 457–465. ACM (2011)

52. Ramtekkar, U.P., Reiersen, A.M., Todorov, A.A., Todd, R.D.: Sex and age differences in attention-deficit/hyperactivity disorder symptoms and diagnoses: implications for dsm-v and icd-11. Journal of the American Academy of Child & Adolescent Psychiatry **49**(3), 217–228 (2010)
53. Ranzato, M., Hinton, G., Lecun, Y.: Guest editorial: Deep learning. International Journal of Computer Vision **113**(1), 1–2 (2015)
54. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents, pp. 37–44. ACM (2010)
55. Rodrigues, R.G., das Dores, R.M., Camilo-Junior, C.G., Rosa, T.C.: Sentihealth-cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. International journal of medical informatics **85**(1), 80–95 (2016)
56. Salas-Zárate, M.d.P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M.Á., Valencia-García, R.: Sentiment analysis on tweets about diabetes: An aspect-level approach. Computational and mathematical methods in medicine **2017** (2017)
57. Sarker, A., Mollá-Aliod, D., Paris, C., et al.: Outcome polarity identification of medical papers (2011)
58. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), e73,791 (2013)
59. Sharif, H., Zaffar, F., Abbasi, A., Zimbra, D.: Detecting adverse drug reactions using a sentiment classification framework (2014)
60. Sidana, S., Mishra, S., Amer-Yahia, S., Clausel, M., Amini, M.R.: Health monitoring on social media over time. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, pp. 849–852. ACM, New York, NY, USA (2016). DOI 10.1145/2911451.2914697. URL `http://doi.acm.org/10.1145/2911451.2914697`
61. Slutske, W.S., Jackson, K.M., Sher, K.J.: The natural history of problem gambling from age 18 to 29. Journal of abnormal psychology **112**(2), 263 (2003)
62. Snyder, P.J., Peachey, H., Hannoush, P., Berlin, J.A., Loh, L., Lenrow, D.A., Holmes, J.H., Dlewati, A., Santanna, J., Rosen, C.J., et al.: Effect of testosterone treatment on body composition and muscle strength in men over 65 years of age. The Journal of Clinical Endocrinology & Metabolism **84**(8), 2647–2653 (1999)
63. Søgaard, A., Goldberg, Y.: Deep multi-task learning with low level tasks supervised at lower layers. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 231–235 (2016)
64. Solovyev, V., Ivanov, V.: Knowledge-driven event extraction in russian: corpus-based linguistic resources. Computational intelligence and neuroscience **2016**, 16 (2016)
65. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association (2003). URL `http://cogprints.org/3164/`
66. Tutubalina, E., Nikolenko, S.: Inferring sentiment-based priors in topic models. In: Mexican International Conference on Artificial Intelligence, pp. 92–104. Springer (2015)
67. Tutubalina, E., Nikolenko, S.: Automated prediction of demographic information from medical user reviews. In: International Conference on Mining Intelligence and Knowledge Exploration, pp. 174–184. Springer (2016)
68. Tutubalina, E., Nikolenko, S.: Demographic prediction based on user reviews about medications. Computación y Sistemas **21**(2), 227–241 (2017)
69. Tutubalina, E., Nikolenko, S.I.: Constructing aspect-based sentiment lexicons with topic modeling. In: Proc. 5th International Conference on Analysis of Images, Social Networks, and Texts, pp. 208–220 (2016)
70. Tutubalina, E., Nikolenko, S.I.: Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. Journal of Healthcare Engineering **2017**, 9451342 (2017)
71. Volkova, S., Van Durme, B.: Inferring user political preferences from streaming communications. In: In Proceedings of the Association for Computational Linguistics (ACL) (2014)

72. Xia, L., Gentile, A.L., Munro, J., Iria, J.: Improving patient opinion mining through multi-step classification. In: TSD, vol. 5729, pp. 70–76. Springer (2009)
73. Yalamanchi, D.: Sideffective-system to mine patient reviews: sentiment analysis. Ph.D. thesis, Rutgers University-Graduate School-New Brunswick (2011)
74. Yang, Z., Kotov, A., Mohan, A., Lu, S.: Parametric and non-parametric user-aware sentiment topic models. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 413–422. ACM (2015)