ISSN 2007-9737

# Demographic Prediction Based on User Reviews about Medications

Elena Tutubalina[1], Sergey Nikolenko[2,3]

[1] Kazan (Volga Region) Federal University, Kazan, Russia

[2] National Research University Higher School of Economics, Laboratory for Internet Studies, St. Petersburg, Russia

[3] Steklov Institute of Mathematics at St. Petersburg, Russia

tlenusik@gmail.com, sergey@logic.pdmi.ras.ru

**Abstract.** Drug reactions can be extracted from user reviews provided on the Web, and processing this information in an automated way 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 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 feature rich classifiers, extensions of topic models, and deep neural networks (both convolutional and recurrent architectures) for this problem.

**Keywords.** Demographic prediction, user reviews, medications.

## 1 Introduction

Modern medical studies increasingly use nonstandard sources of information to obtain new data related to medical conditions, efficiency of drugs, their adverse effects, interactions between different drugs, and so on. One such source of information can be provided by the drug users themselves, in the form of free-text web reviews, social media posts, and other user-generated texts. These sources have been successfully used, for instance, to monitor adverse drug reactions (ADRs), making it possible to detect rare and underestimated ADRs through the users complaining about their health on social networks or specialized forums [36].

However, it may be important for the medical field to learn more than just the existence of an adverse reaction from a text review. Drugs may exhibit different behavior on people with different age, gender, or other parameters that will often be unknown for a text scraped from an Internet forum. Hence, the problem arises to mine demographic information from free-text medical 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 databases of medical reviews from health-related Web sites with user-generated content, namely *WebMD* and *AskaPatient*, and have trained models to predict the age and gender of users who wrote these reviews. We propose a classification approach based on a classical classifier (we compare SVM and Maximum Entropy classifier, i.e., logistic regression) which is augmented by sets of features based on recently developed novel approaches to text mining: topic models, including the Partially Labeled Topic Model, and features based on word embeddings. We show that the resulting classifier performs significantly better than the baseline.

This work is a significantly extended journal version of the paper [40]; compared to the conference version, we have changed the approach to

baseline classifiers, making them into feature-rich classifiers with topics and word embeddings as features. We have also significantly extended the set of said features, adding new domain-specific information to aid the classifiers. Therefore, the experimental part of this work is new compared to [40], and the results have been substantially improved.

The paper is organized as follows. In Section 2, we survey related work on mining drug-related information from social media and other user-generated texts. Section 3.1 defines models for information extraction from text that we compare in this work: we present the features and briefly introduce topic models with user attributes and distributed word representations. We present experimental results in Section 4 and conclude with Section 5.

## 2 Related Work

The use of social media for medical and pharmacological data mining has been on the uprising since early 2010s; the term "pharmacovigilance" has been coined for automated monitoring of social media for potentially adverse drug effects and interactions; see also media articles about these effects [14, 37]. One of the first works on this subject [13] analyzed user posts regarding six drugs from a health-related social network. A comprehensive review of text mining techniques as applied to drug reaction detection can be found in [9]. We also note a Social Media Mining Shared Task Workshop (organized as part of the Pacific Symp. on Biocomputing 2016) devoted to mining pharmacological and medical information from social media, with a competition based on a published dataset [35].

In [6], authors identify ADRs from texts on health-related online forums. They used dictionary-based drug detection, extracting symptoms with a combination of dictionary-based and pattern-based methods. A lift measure (also known as pointwise mutual information) was computed to evaluate the likelihood of drug-ADR relation and chi-square test was used to evaluate the statistical significance of the lift measure. Several case studies of drugs showed that some ADRs were reported prior to FDA approval. One limitation of this work is the number of annotated examples in test data: less than $500$ ADRs for evaluation.

In [32], existing machine learning dictionary-based approaches were used to identify disease names from user reviews about top 180 most frequently searched medications on the forum WebMD, using a rule-based system to extract beneficial effects of the drug. In order to identify candidates for drug repurposing, authors removed known drug indications and manually reviewed the comments without FDA reports. The main limitation of this work is the lack of an annotated corpus to evaluate the proposed method. The work [42] shows an experiment for ten drugs and five ADRs to examine associations between them on texts from online healthcare communities using association mining techniques. The FDA alerts served as a gold standard to evaluate the associations discovered between drugs and ADRs. We also note a series of works specifically on Spanish language social media [15, 36].

Usually, pharmacovigilance studies employ simple classifiers to extract information on drug effects or interactions. For example, to mine drug-related information from a stream of *Twitter* data, a recent work [24] uses a cascade of simple input filters followed by an SVM classifier, reporting good discovery results, while [44] proposes a weighted average ensemble of four classifiers: one based on a handmade lexicon, two on $n$-grams, and one on word embeddings.

On the other hand, drug testing and discovery of drug effects and interactions requires one to know demographic information about a user since drug effects can differ significantly depending on the user. This leads to the need to mine demographic information about the authors together with the user-generated texts themselves. When such information is provided, e.g., when the texts are collected from *facebook* users 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* or *AskaPatient*; often demographic information can be known for a minority of users but not all. Hence, the problem

arises to predict user demography based on the texts of user reviews.

In natural language processing, predicting demographic features based on free text falls into a large classical field of authorship analysis, attribution and author verification studies [12, 45]; we refer to surveys [3, 38, 39] for details and references. Numerous works on the topic have been published based on the results of the shared Author Profiling Tasks at digital text forensics events by PAN initiative [2, 5, 7, 27–30]. However, authorship analysis seldom extends to medical issues: for example, the work [23] attempts to screen Twitter users for depression based on their tweets, but to the best of our knowledge, previous work has not attempted to automatically mine demographic information unless it was provided explicitly. In this work, we begin to fill this gap, providing first results on automated predictions of demographic based specifically on medical reviews.

## 3 Classification Methods

### 3.1 Models

In this section, we describe two different approaches for demographic prediction applied to a collection of user comments about medications. First, we describe our feature-rich machine learning classifiers. Second, we describe neural networks that rely on word representations learned from unannotated reviews.

### 3.2 Basic Classifiers and their Features

We formulate the prediction of user attributes as a classification problem. In order to perform the classification, we apply two supervised approaches with a set of hand-crafted features:

(1) support vector machine (SVM);

(2) logistic regression, also called the Maximum Entropy classifier (MaxEnt).

These approaches have been known to achieve the best results in various classification tasks, including sentiment and subjectivity classification [11, 41], ADR classification [34], and demographic prediction [22, 31]. Our classifiers leverages a variety of surface-form, semantic, cluster-based, distributed and lexicon features described below.

The entire set of features used in our classifiers consists of the following subsets:

— **Word ngrams** (NGR): occurrence of contiguous sequences of 1, 2, and 3 tokens; the maximum number of features are 25,000;

— **Drug classification groups** (ATC): drug names are classified in groups at five different levels using the DrugBank database and the ATC classification system;

— **Automatically generated lexicons** (PMI): for each token occurring in a text and present in our automatic lexicon, we use its score to compute the number of tokens with $\mathrm{score}(w) > 0$ and sum of these scores, the number of tokens with $\mathrm{score}(w) < 0$ and sum of these scores, the total score, and maximal and minimum scores; all scores and sums are averaged for each review;

— **Sentiment lexicons** (SENT): for each of the sentiment lexicons (Bing Lius Lexicon and MPQA Subjectivity Lexicon), we compute the following two features: average sum of positive scores for the tokens and average sum of negative scores for the tokens;

— **ADR lexicon** (ADR): presence/absence of ADR mentions using the lexicon;

— **Clusters** (CL): presence/absence of tokens from each of the 150 clusters;

— **Topics** (TPC): presence/absence of tokens from each of 150 topics;

— **Word embeddings** (EMB): the real-valued vector of each word as described in Section 3.4.

In the remainder of this subsection, we define each of these items in detail.

**ATC classification.** In the Anatomical Therapeutic Chemical (ATC) classification system, biomedical and chemical entities are divided into different groups according to the organ on which they act and their therapeutic, pharmacological, and chemical properties. Using the DrugBank database, we find the presence of a drug in each class up to 5 levels. For example, Prozac (Fluoxetine) is associated with the ATC code N06AB03 and classified into this code and the following codes from higher levels: 'elective serotonin reuptake inhibitors' (N06AB), 'antidepressants' (N06A), 'psychoanaleptics' (N06), 'nervous system' (N). We use these features to incorporate domain-specific medical knowledge into the classification process.

**Automatically generated lexicon.** The key idea of this automatically generated lexicon is to take advantage of a large corpus of weakly labeled texts, where authors assign several predefined labels to each text. Following state-of-art approaches for sentiment analysis [11], we automatically generated a lexicon based on the score for each token (w) (with frequency greater or equal than 10) in the Health dataset:

$$\text{score}(w) = \text{PMI}(w, cat) - \text{PMI}(w, oth),$$

where

$$\text{PMI}(w, cat) = \log \frac{p(w, cat)}{p(w) * p(cat)}$$

is the pointwise mutual information, $cat$ denotes all texts associated with the particular category, $oth$ denotes all texts in other categories, and $p(w, pt)$ are probabilities of $w$ occurring in the texts labeled with a particular category. As categories we separately use age and gender attributes.

**Sentiment lexicons.** We used Bing Lius Lexicon[1] and the MPQA Subjectivity Lexicon[2]. We assign the score of +1 for positive entries and the score of -1 for negative entries from the Bing Lius Lexicon. For the MPQA Subjectivity Lexicon, we assign scores +0.5/-0.5 and +1/-1 for weak and strong associations respectively.

[1] http://www.cs.uic.edu/ liub/FBS/opinion-lexicon-English.rar
[2] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

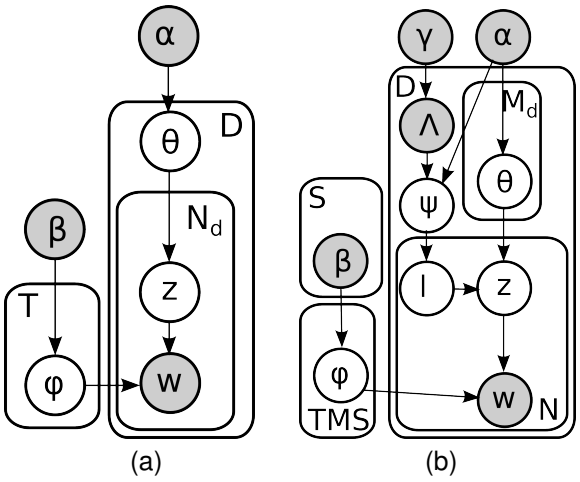


**Fig. 1.** Probabilistic graphical topic models: (a) the basic LDA model; (b) PLDA

**ADR lexicon.** We assume that patients experience different adverse drug reactions that may depend on age and gender. In order to use medical information specific to demographic groups, we develop the exact lookup based on ADR lexicon from the paper [34]. The lexicon contains 16,183 ADRs from several resources: the COSTART vocabulary created by the FDA for post-market surveillance of ADRs, the SIDER side effect resource, and the Canada Drug Adverse Reaction Database, SIDER II and the Consumer Health Vocabulary.

**Cluster-based features.** Clusters reduce the sparsity of the token space as an alternative representation of text. We use the Brown algorithm, i.e., a hierarchical clustering algorithm [4]. The algorithm partitioned the words into a set of 150 clusters, and we add features corresponding to the presence or absence of specific clusters in the review.

Next, we discuss the last two classes of features that come from topic models and word embeddings respectively.

**Table 1.** Sample topics discovered by PLDA for the tag ”female” and ”male”

| # | topic words |
|---|---|
| | **male** |
| 1 | muscle left pain legs hands joint neck feet pains burning arms aches body ingling walk |
| 2 | effect sexual longer however difficult positive side sex negative control reduced libido |
| 3 | stomach diarrhea food eat acid cramps nexium gas upset nausea reflux pains |
| 4 | infection throat days rash itching reaction sinus body nose cough face fever |
| 5 | meds wife make gave finally times god home big people care end rest things house stay |
| | **female** |
| 1 | stomach nausea diarrhea eat food cramps upset sick vomiting acid bloating gas constipation |
| 2 | body hands rash feet reaction legs itching face swelling arms burning tingling allergic swollen |
| 3 | days infection throat prescribed sinus sore cough headache nose antibiotic fever ear |
| 4 | feel things happy person family dont anymore husband care longer crying depressed job |

### 3.3 Topic Models

For topic-based features, we employ the latent Dirichlet allocation (LDA) model, a classical topic model. 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 probabilistic graphical model of basic LDA is shown on Fig. 1a. The model introduces Dirichlet priors with parameters $\alpha$ for topic vectors $\theta$, $\theta \sim \mathrm{Dir}(\alpha)$, and $\beta$ for word distributions $\phi$, $\phi \sim \mathrm{Dir}(\beta)$ (we assume the Dirichlet priors are symmetric, as they usually are). A document is generated word by word: for each word, first sample its topic index $t$ from $\theta_d$, $t \sim \mathrm{Mult}(\theta_d)$, then sample the word $w$ from $\phi_t$, $w \sim \mathrm{Mult}(\phi_t)$. 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 $w$). In the basic LDA model, inference proceeds with *collapsed Gibbs sampling*, where $\theta$ and $\phi$ 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 $z_j$. Samples are then used to estimate model variables:

$$\theta_{td} = \frac{n_{w,t,d} + \alpha}{n_{w,*,d} + T\alpha}, \quad \phi_{wt} = \frac{n_{w,t,*} + \beta}{n_{*,t,*} + W\beta}.$$

We also experimented with *Partially Labeled Topic Model* (PLDA) [26]. PLDA incorporates user meta-data tags (e.g., location, gender, or age) together with topics. 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, and words are conditioned on the latent topics and tags. The probabilistic graphical model of PLDA is shown on Fig. 1b. The Gibbs sampling step proceeds as

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

An important characteristic feature of topic models is that they can be mined for qualitative results that are easy to interpret and can validate their performance. For example, Table 1 shows topics discovered by the PLDA model based on a unigram representation of reviews related to each gender; note that the distinction between "male" and "female" topics does indeed reflect common medical knowledge.

### 3.4 Distributed Word Representations

The other class of models in our study is very different in nature from topic models. We compare results produced by topic models with classification models based on *word2vec* embeddings processed by recurrent and convolutional neural networks (RNNs and CNNs).

Recent advances in distributed word representations have made it into a method of choice for modern natural language processing [8]. Distributed word representations are models that map each word occurring in the dictionary to a Euclidean space, attempting to capture semantic relationships between the words as geometric relationships in the Euclidean space. In a classical word embedding model, one first constructs a vocabulary with one-hot representations of individual words, where each word corresponds to its own dimension, and then trains representations for individual words starting from there, basically as a dimensionality reduction problem. For this purpose, researchers have usually employed a model with one hidden layer that attempts to predict the next word based on a window of several preceding words. Then representations learned at the hidden layer are taken to be the word's features.

The *word2vec* embeddings come in two flavors, both introduced in [16]: *Continuous Bag-of-Words* (CBOW) and *skip-gram*. During its learning, a CBOW model is trying to reconstruct the words from their contexts with a network whose architecture is shown on Fig. 2a; the training process for this model proceeds as follows:

(1) each of the inputs of this network is a one-hot encoded vector of size $|V|$, where $V$ is the vocabulary;

(2) when computing the output of the hidden layer, we take an average of all input vectors; the hidden layer is basically a matrix of vector embeddings of words, so the $n$th row represents an embedding of the $n$th word in the vocabulary;

(3) the output layer represents a score $u_j$ for each word in the vocabulary; to obtain the posterior, which is a multinomial distribution, we then use the softmax

$$\hat{P}(w_t|w_1^{t-1}) = \frac{\exp(u_j)}{\sum_{j'=1}^{t}\exp(u_{j'})},$$

so the loss function is

$$E = -\log p(w_t|w_1^{t-1}) = -u_j + \log\sum_{j'=1}^{|V|}\exp(u_{j'}).$$

The skip-gram model operates inversely, predicting the context from the word, which can be seen from its network architecture shown on Figure 2b. Here the target is an input word, and the output layer, in turn, now represents $C$ multinomial disctibutions

$$\hat{P}(w_1^{t-1}|w_t) = \frac{\exp(u_{cj})}{\sum_{j'=1}^{i}\exp(u_{j'})}$$

with the loss computed as

$$E = -\log p(w_1^{t-1}|w_t) = \\ = -\sum_{c=1}^{C} u_{jc} + C\log\sum_{j'=1}^{|V|}\exp(u_{j'}).$$

The idea of word embeddings has been applied back to language modeling in [17, 18, 21], and starting from the works of Mikolov et al. [16, 19], word representations have been used for numerous NLP problems, including text classification, extraction of sentiment lexicons, part-of-speech tagging, syntactic parsing, and others.

Word embedding models represent each word using a single real-valued vector. Such representation groups together words that are semantically and syntactically similar [20]. We used word2vec from Gensim library to train embeddings on the

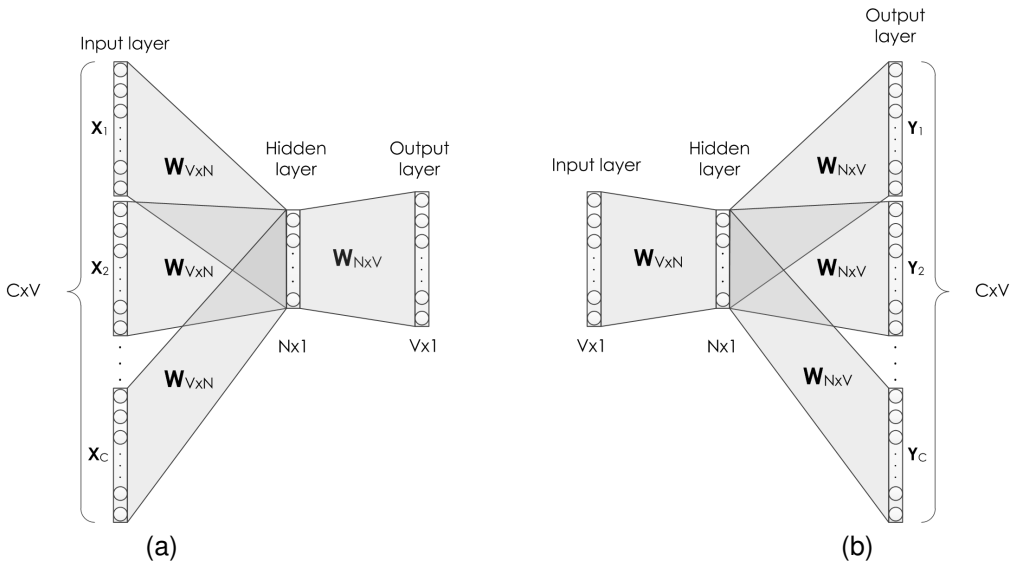


**Fig. 2.** Illustration of the *word2vec* models: (a) CBOW, (b) skip-gram [16, 33]

Health Dataset. We applied Continuous Bag of Words model with the following parameters: vector size of 200, the length of local context of 10, negative sampling of 5, vocabulary cutoff of 10. Below, we refer to our pre-trained vectors as **HealthVec**. We also experimented with another published word vector PubMedVec (2,351,706 terms) trained on biomedical literature indexed in PubMed [25]. PubMed comprises more than 26 million citations for biomedical literature from bibliographic database MEDLINE, journal articles, life science journals, and online books.

### 3.5 Neural Network Classifiers

In this work, we compare two modern approaches to natural language processing with neural networks: traditional recurrent architectures, specifically LSTM-based recurrent networks, and convolutional neural networks (CNN). In the recurrent part, we use an architecture with multiple LSTM layers, where higher layers use the sequence of outputs from the previous layer of LSTMs, and on the top level the LSTM outputs are combined into the final layer which does the actual prediction.

While CNNs have been most successfully used for image processing, recent applications of CNNs to natural language processing also produce state of the art results. In an NLP task, convolutional layers are still interleaved with subsampling max-pooling layers, but this time the convolutions are one-dimensional rather than two- or three-dimensional as in images and video. Here, we use a convolutional model similar to the one recently presented in [10] for semantic sentence classification; this model has the following characteristic features:

— it is not as deep as computer vision models and involves only one convolutional layer with max-over-time pooling and a softmax output;

— regularization is achieved through dropout; the authors report a 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.

## 4 Evaluation

### 4.1 Datasets

For experimental evaluation, we have crawled health-related reviews from two health hotel review sites: (i) *WebMD*[3] and (ii) *AskaPatient*[4].

*WebMD* is a health information services website that aims to provide objective, trustworthy, and valuable health information. We have crawled 217,485 reviews from authors tagged as "Patient". Each review contains the following fields: 1) date when the review was written, 2) condition for taking treatment, 3) free-text review given for the effects caused due to the use of the drug, 4) user attributes such as gender and age. Gender tags are "Male" or "Female", and predefined age tags in the dataset are "19–24", "25–34", "35-44", "45-54", "55-64", "65-74", or "75 or over". In this study, we combine some of the age tags and divide user attributes into three major age groups: "19-34", "35-64", and "65 and over".

*AskaPatient*[5] website aims to empower patients by allowing them to share and compare their medical experiences. We have crawled $113{,}093$ reviews. Since users often confuse two free-text fields about a drug, we have concatenated the "side effects" and "comments" fields, treating the result as a full review. Similar to *WebMD*, reviews from *AskaPatient* contain textual information, reason for taking treatment and user attributes (without predefined list of age groups).

In contrast with our previous work, we split our corpora into training and testing parts further referred as **WebMD** and **AskaPatient** (used by the ML and DL algorithms) and a free-text corpus of in-domain texts called the **Health dataset** (used to compute PMI, topics, and word representations). In order to create robust methods and exclude drugs with highly imbalanced genders (e.g., birth control pills), we use reviews associated with 5 most commented conditions for training/testing. For WebMD, review authors select a condition from a predefined list for every drug. For AskaPatient, the "reason" is a free-text field.

Table 2 summarizes the statistics of both datasets used in our study. The WebMD dataset contains 20,693 reviews with the age group "35-64", 7,410 reviews with the age group "19-34", and 7,519 reviews with the age group "65 and over". The total numbers of tokens in the WebMD and AskaPatient datasets are 2,818,429 and 1,051,969, respectively. The total numbers of unique tokens in the WebMD and AskaPatient datasets are 33,411 and 18,825, respectively.

### 4.2 Model Parameters

In order to get local features from a review with CNNs we have used multiple filters of different lengths [10]. We separated out 10% of the training dataset to form the validation set which was used to evaluate different model parameters. We used a sliding max-pooling window of size $2$ to get features through filters. Pooled features are then fed to a fully connected feed-forward neural network (with dimension $100$) which uses rectified linear units as output activations. Then we apply a softmax classifier with the number of outputs equal to the number of classes. We applied dropout rate of $0.5$ to the fully connected layer and trained the network for $20$ epochs; on the other hand, we did not apply dropout after the embedding layer since in our experiments this led to lower results achieved by CNNs on the validation set. The width of the convolution filters is set to $3$, each with $100$ filters. Additionally, we employ early stopping after two epochs with no improvement on the validation set. Embedding layers are not trainable for all networks. We set mini-batch size to $256$ and $128$ for the WebMD and AskaPatient datasets, respectively.

In our experiments with recurrent neural networks, we used a standard GRU or LSTM architecture on top of the embedding layer that implemented pre-trained word embeddings. Similar to [1], the resulting sequence of vectors serves as the input to the network. We experimented with shallow GRU/LSTM, two-layer GRU, and used $100$ units on each layer with the Adam optimizer and rectified linear units as output

[3] http://www.webmd.com
[4] http://www.askapatient.com
[5] http://www.askapatient.com

**Table 2.** Summary statistics for the experimental datasets; number of reviews with a given label in shown in parentheses

| Dataset | Top-5 Conditions | Gender | Age groups |
|---|---|---|---|
| **WebMD** | high blood pressure (10201)<br>pain (9306)<br>depression (7340)<br>chronic trouble sleeping (3454)<br>attention deficit disorder with hyperactivity (3021) | Female (23343)<br>Male (9979) | 45-54 (8430)<br>55-64 (7056)<br>35-44 (6207)<br>19-34 (7410)<br>65 or over (4219) |
| **AskaPatient** | depression (3170)<br>anxiety (1603)<br>uti (1545)<br>insomnia (1329)<br>high blood pressure (1270) | Female (6356)<br>Male (2561) | |

activation. Similar to CNN, GRU layer is then fed to a fully connected feed-forward neural network (with dimension $100$). Other parameters are adopted from CNN settings.

We tested and compared the following vectors:

— **NewsVec**: commonly used word embeddings *GoogleNews-vectors-negative300*[6] trained on part of Google News dataset (about 100 billion words).

— **PubmedVec**: word vectors trained on biomedical scientific literature *PubMed* [25];

— **HealthVec**: word vectors trained on product reviews from the Health dataset.

The general statistics are presented in Table 3. We also observed better classification results after normalizing each vector by dividing it by its 2-norm.

For SVM and MaxEnt classifiers, we used LinearSVC and LogisticRegression with default parameters from the NLTK library[7]. We used Liang's implementation of the Brown hierarchical word clustering algorithm[8].

We used the Mallet[9] library to generate topics. The number of sampling iterations was set to 1000. We used default hyperparameters, took top 20 words for each topic, and evaluated 50, 100, and 150 topics on the validation set. The best results were achieved with 150 topics. We also implemented PLDA for comparison, adopting its parameters from LDA. For further evaluation, we selected topics from LDA rather than PLDA since they produced better results on validation data.

**Table 3.** Statistics of *word2vec* embeddings

| Embeddings | Dimension | # of tokens |
|---|---|---|
| GoogleNews | 300 | 3,000,000 |
| PubMed | 200 | 2,351,706 |
| HealthVec | 200 | 31,482 |

### 4.3 Results

In this section, we describe our experiments with feature-rich classifiers and deep learning models. We performed pre-processing by lower-casing all words. We performed $5$-fold cross-validation and computed precision (P), recall (R), and F1-measure (F1), showing the macro-averaged results in Table 4 (gender prediction) and Table 5 (age prediction). The tables also show the best results for each model type in every column highlighted in bold.

The main result, which might look surprising at first, is that standard classifiers, when enriched with a large number of various features, outperform even the best neural network approaches that we have been able to train. Specifically, CNNs and RNNs are able to achieve better precision than SVM and MaxEnt but lose significantly in recall and therefore in the aggregate F1 measure. Moreover, Tables 4 and 5 show the variances

[6] https://code.google.com/archive/p/word2vec/
[7] http://www.nltk.org
[8] https://github.com/percyliang/brown-cluster
[9] http://mallet.cs.umass.edu

**Table 4.** Gender prediction (macro-averaged, 2 classes)

| Model and features | WebMD | | | AskaPatient | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| SVM classifier (first column shows feature set) | | | | | | |
| NGR (1-, 2-, and 3-grams) | 0.645 | 0.649 | 0.647±0.006 | 0.651 | 0.646 | 0.648±0.012 |
| NGR+ATC | 0.647 | 0.651 | 0.649±0.006 | 0.658 | 0.652 | 0.655±0.016 |
| NGR+EMB+TCS+CL | 0.651 | 0.653 | 0.655±0.007 | 0.650 | 0.650 | 0.650±0.009 |
| NGR+EMB+TCS+CL+SENT+PMI+ADR | **0.674** | **0.676** | **0.675±0.003** | **0.665** | **0.657** | **0.660 ±0.009** |
| MaxEnt classifier (first column shows feature set) | | | | | | |
| NGR | 0.671 | 0.662 | 0.666±0.005 | 0.675 | 0.653 | 0.660±0.008 |
| NGR+ATC | 0.674 | 0.664 | 0.668±0.006 | 0.679 | 0.657 | 0.665±0.013 |
| NGR+EMB+TCS+CL | 0.676 | 0.670 | 0.673±0.006 | 0.670 | 0.661 | 0.664±0.010 |
| NGR+EMB+TCS+CL+SENT+PMI+ADR | **0.702** | **0.691** | **0.695±0.006** | **0.683** | **0.665** | **0.672±0.009** |
| Neural networks for end-to-end classification | | | | | | |
| CNN, HealthVec, $[1, 2, 3]$ filters | 0.706 | 0.651 | 0.663±0.011 | 0.684 | 0.635 | **0.643±0.027** |
| CNN, HealthVec, $[1, 2, 3]$, trainable emb. | 0.678 | 0.649 | 0.657±0.007 | 0.655 | **0.636** | 0.642±0.011 |
| CNN, HealthVec, $[1, 2, 3, 4]$ filters | 0.702 | 0.653 | 0.658±0.007 | 0.684 | 0.614 | 0.620±0.012 |
| CNN, HealthVec, $[1, 2, 3, 4]$, trainable emb. | 0.674 | **0.664** | **0.668±0.007** | 0.673 | 0.634 | 0.642±0.021 |
| CNN, NewsVec, $[1, 2, 3]$ filters | **0.707** | 0.645 | 0.653±0.029 | **0.701** | 0.610 | 0.612±0.048 |
| CNN, PubmedVec, $[1, 2, 3]$ filters | 0.705 | 0.637 | 0.646±0.021 | 0.635 | 0.607 | 0.614±0.055 |
| 2-layer GRU, HealthVec | 0.674 | 0.648 | 0.654±0.019 | 0.622 | 0.588 | 0.590±0.024 |
| 1-layer GRU, HealthVec | 0.680 | 0.632 | 0.640±0.015 | 0.599 | 0.545 | 0.523±0.077 |

**Table 5.** Age prediction (macro-averaged, 3 classes)

| Model and features | WebMD | | |
|---|---|---|---|
| | P | R | F1 |
| SVM classifier (first column shows feature set) | | | |
| NGR (1-, 2-, and 3-grams) | 0.514 | 0.513 | 0.513±0.011 |
| NGR+ATC | 0.526 | 0.524 | 0.525±0.002 |
| NGR+EMB+TCS+CL | 0.516 | 0.518 | 0.517±0.005 |
| NGR+EMB+TCS+CL+SENT+PMI+ADR | **0.540** | **0.539** | **0.539±0.007** |
| MaxEnt classifier (first column shows feature set) | | | |
| NGR | 0.562 | 0.521 | 0.536±0.004 |
| NGR+ATC | 0.566 | 0.527 | 0.542±0.003 |
| NGR+EMB+TCS+CL | 0.560 | 0.529 | 0.541±0.004 |
| NGR+EMB+TCS+CL+SENT+PMI+ADR | **0.574** | **0.544** | **0.557±0.008** |
| Neural networks for end-to-end classification | | | |
| CNN, HealthVec, $[1, 2, 3]$ filters | 0.615 | 0.490 | 0.510±0.014 |
| CNN, HealthVec, $[1, 2, 3]$, trainable emb. | 0.585 | 0.512 | 0.532±0.013 |
| CNN, HealthVec, $[1, 2, 3, 4]$ filters | 0.637 | 0.482 | 0.504±0.012 |
| CNN, HealthVec, $[1, 2, 3, 4]$, trainable emb. | 0.588 | **0.514** | **0.536±0.006** |
| CNN, PubmedVec, $[1, 2, 3]$ filters | **0.648** | 0.467 | 0.488±0.009 |
| 2-layer GRU, HealthVec | 0.618 | 0.485 | 0.483±0.017 |
| 1-layer GRU, HealthVec | 0.530 | 0.409 | 0.396±0.046 |

**Table 6.** Representative MaxEnt features for male and female patients with different conditions (WebMD dataset)

| **Pain** | | | | **High blood pressure** | | | |
|---|---|---|---|---|---|---|---|
| **Female** | | **Male** | | **Female** | | **Male** | |
| for fibromyalgia | 1.069 | my wife | 2.027 | my hair | 1.904 | old male | 1.660 |
| fibromyalgia pain | 0.937 | for shoulder | 0.976 | hair loss | 1.351 | my wife | 1.478 |
| my migraines | 0.918 | back fusion | 0.814 | have gained | 1.217 | lower my blood | 0.991 |
| for arthritis | 0.902 | my knee pain | 0.693 | so tired | 1.066 | sex drive | 0.910 |
| muscle relaxer | 0.848 | sleep at night | 0.658 | terrible cough | 0.925 | sexual desire | 0.841 |
| severe migraine | 0.747 | pain level | 0.651 | swollen ankles | 0.875 | erectile dysfunction | 0.811 |
| old female | 0.715 | scar tissue | 0.643 | leg cramps | 0.863 | frequent urination | 0.662 |
| my headaches | 0.700 | kidney stones | 0.578 | hot flashes | 0.859 | heart attack | 0.585 |
| throwing up | 0.688 | chronic knee | 0.578 | severe headaches | 0.706 | my kidney | 0.557 |
| allergic to | 0.676 | bulging disk | 0.564 | muscle pain | 0.676 | ankle swelling | 0.547 |
| **Depression** | | | | **Attention deficit disorder with hyperactivity** | | | |
| **Female** | | **Male** | | **Female** | | **Male** | |
| my husband | 1.281 | my wife | 2.723 | my daughter | 0.706 | my wife | 1.580 |
| loss of appetite | 0.741 | some sexual | 0.836 | with adhd | 0.675 | old male | 1.042 |
| gained weight | 0.739 | my girlfriend | 0.777 | my son | 0.548 | very effective | 0.848 |
| very happy | 0.676 | anxiety disorder | 0.758 | weight loss | 0.522 | to urinate | 0.661 |
| my kids | 0.656 | alcohol and | 0.705 | my child | 0.521 | my brain | 0.653 |
| crying spells | 0.642 | an erection | 0.662 | to help | 0.517 | abdominal pain | 0.478 |
| lost weight | 0.599 | my marriage | 0.600 | my mood | 0.445 | over the years | 0.414 |
| hot flashes | 0.550 | diet and exercise | 0.508 | my heart | 0.412 | personal relationships | 0.360 |
| my moods | 0.510 | sexual dysfunction | 0.504 | my husband | 0.391 | an alcoholic | 0.324 |
| my daughter | 0.507 | lack of appetite | 0.499 | old female | 0.339 | my girlfriend | 0.324 |

**Table 7.** Representative features obtained by MaxEnt for different age groups (WebMD dataset)

| **High blood pressure** | | | | | |
|---|---|---|---|---|---|
| **19-34** | | **35-64** | | **65 and over** | |
| birth control | 0.806 | light headed | 0.876 | was normal | 1.250 |
| during pregnancy | 0.707 | lowered bp | 0.820 | too expensive | 1.216 |
| chest pains | 0.667 | frequent urination | 0.811 | feet and ankles | 0.817 |
| my pregnancy | 0.657 | for high bp | 0.801 | dry eyes | 0.814 |
| my headaches | 0.625 | my sleep | 0.768 | the price | 0.813 |
| low dosage | 0.589 | feeling tired | 0.703 | breathing problem | 0.809 |
| for my blood | 0.588 | heartburn and | 0.685 | my cardiologist | 0.779 |
| get pregnant | 0.557 | sex drive | 0.671 | life threatening | 0.691 |
| muscle cramps | 0.524 | rapid heart rate | 0.629 | vision problems | 0.626 |
| extreme fatigue | 0.491 | my insurance | 0.546 | my sodium | 0.618 |

of the F1-measure in our cross-validation results, indicating that the advantage of SVM and MaxEnt in F1-measure is statistically significant.

This seemingly unexpected result is, in our opinion, due to two main reasons. First, we are free to augment standard classifiers with any features we want, thus using a wide variety of external information that is unavailable to the neural networks, which have to rely on text only. It is unclear how to introduce all of the features that we used for SVM and MaxEnt into the neural networks, and it would require a separate complex study, both theoretical and practical, to incorporate these features.

Second, the dataset size in this case is probably not large enough for the neural networks to shine. Since we used suitable regularization we did not experience strong overfitting in the neural

networks, but general rules of thumb suggest that our supervised datasets are too small for the expressive power of complex neural networks to have significant effect.

Thus, our results suggest that while neural network approaches often define the state of the art in modern natural language processing, in problems where rich additional information can be made available, especially in domain-specific problems with well defined domains (such as medicine in this case), classical machine learning approaches can still be very useful and can still be successfully used in practical settings.

Secondary results include two conclusions from Tables 4 and 5. First, while adding more features is usually obviously beneficial, this did not hold for ATC features in our experiments: they helped much less than others and even deteriorated the results. This is probably due to the fact that a relatively small dataset size combined with high dimension of ATC features led to overfitting. Second, note that the best results with neural networks are usually obtained in variations where the word embeddings are also trainable. Regardless of the dataset size (which would be much too small to train embeddings properly), in our experience making embeddings trainable (i.e., slightly fitting them in the end-to-end supervised network starting from unsupervised vectors) appears to be beneficial almost always and should be adopted in most settings.

We have also performed qualitative analysis of our results. In particular, we have extracted and analysed the most representative $n$-grams for various conditions. Tables 6 and 7 present the most representative features (excluding numeric features) for one gender over another and for a certain age group over other age classes, respectively. For this experiments, we used the MaxEnt classifier trained on the set of 2- and 3-grams extracted from the review texts. The tables indicate that key terms change with age or gender, reflecting quite natural progressions that match well with medical and commonsense intuition. Hence, our classifiers can also be used to mine qualitative information from a dataset of medical reviews, perhaps uncovering new common conditions or important factors in a certain user group.

## 5 Conclusion

In this work, we have presented the first results on the practically important problem of automatically learning demographic user features from his or her reviews concerning medical products or services.

We have compared several different models for gender classification and age prediction: baseline classifiers that operate on words and bigrams, feature-rich classifiers with additional information from topic models and word embeddings as well as domain-specific medical information, and convolutional and recurrent neural networks based on *word2vec* embeddings.

Results of our experiments suggest that in settings with relatively small datasets and available external information classical machine learning techniques can outperform neural network approaches. This is due to both dataset size and the fact that while it is hard to tailor neural networks to a specific form of external information, standard classifiers incorporate such new features trivially. We believe that this sample application shows that there is still a place for domain-specific machine learning solutions, especially for relatively small supervised datasets.

## Acknowledgements

This work was supported by the Russian Science Foundation grant no. 15-11-10019. The authors are also sincerely grateful to the anonymous referees whose comments have led to significant improvements in the paper.

## References

1. **Arkhipenko, K., Kozlov, I., Trofimovich, J., Skorniakov, K., Gomzin, A., & Turdakov, D.** (**2016**). Comparison of neural network architectures for sentiment analysis of Russian tweets. *Proceedings of International Conference, Computational Linguistics and Intellectual Technologies*.

2. **Bayot, R. & Gonçalves, T.** (**2016**). Author profiling using SVMs and word embedding averages – notebook for PAN at CLEF. *CLEF Evaluation Labs and Workshop – Working Notes Papers*, Évora, Portugal.

3. **Bouanani, S.E.M.E. & Kassou, I.** (**2014**). Authorship analysis studies: A survey. *International Journal of Computer Applications*, Vol. 86, No. 12, pp. 22–29.

4. **Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., & Lai, J.C.** (**1992**). Class-based n-gram models of natural language. *Computational linguistics*, Vol. 18, No. 4, pp. 467–479.

5. **Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., & Nissim, M.** (**2016**). Gronup: Groningen user profiling – Notebook for PAN. *CLEF 2016, Evaluation Labs and Workshop – Working Notes Papers*, pp. 5–8, Évora, Portugal.

6. **Feldman, R., Netzer, O., Peretz, A., & Rosenfeld, B.** (**2015**). Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, ACM, New York, USA, pp. 1779–1788.

7. **Forner, P., Navigli, R., & Tufis, D. (2013).** *CLEF evaluation labs and workshop – Working notes papers*. pp. 23–26, Valencia, Spain.

8. **Goldberg, Y.** (**2015**). A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726.

9. **Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C.** (**2015**). Text and data mining techniques in adverse drug reaction detection. *ACM Comput. Surv.*, Vol. 47, No. 4, 56:1–56:39.

10. **Kim, Y.** (**2014**). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

11. **Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S.** (**2014**). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 437–442.

12. **Koppel, M., Schler, J., Argamon, S., & Messeri, E.** (**2006**). *Authorship attribution with thousands of candidate*.

13. **Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., & Gonzalez, G.** (**2010**). Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. *Proceedings of the Workshop on Biomedical Natural Language Processing*, BioNLP '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 117–125.

14. **Marcus, A. D.** (**2014**). Researchers fret as social media lift veil on drug trials. *Wall Street Journal*.

15. **Martınez, P., Martınez, J. L., Segura-Bedmar, I., Moreno-Schneider, J., Luna, A., & Revert, R.** (**2016**). Turning user generated health-related content into actionable knowledge through text analytics services. *Computers in Industry*, Vol. 78, pp. 43–56.

16. **Mikolov, T., Chen, K., Corrado, G., & Dean, J.** (**2013**). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

17. **Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S.** (**2010**). Recurrent neural network based language model, *INTERSPEECH*, Vol. 2, No. 3.

18. **Mikolov, T., Kombrink, S., Burget, L., Cernockỳ, J. H., & Khudanpur, S.** (**2011**). Extensions of recurrent neural network language model. *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 5528–5531.

19. **Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J.** (**2013**). Distributed

representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

20. **Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.** (**2013**). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111–3119.

21. **Mnih, A. & Hinton, G. E.** (**2009**). A scalable hierarchical distributed language model. *Advances in neural information processing systems*, pp. 1081–1088.

22. **Nguyen, D., Smith, N. A., & Rosé, C. P.** (**2011**). Author age prediction from text using linear regression. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, pp. 115–123.

23. **Pedersen, T.** (**2015**). Screening twitter users for depression and ptsd with lexical decision lists. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Association for Computational Linguistics, Denver, Colorado, pp. 46–53.

24. **Plachouras, V., Leidner, J. L., & Garrow, A. G.** (**2016**). Quantifying self-reported adverse drug events on twitter: Signal and topic analysis. *Proceedings of the 7th 2016 International Conference on Social Media & Society*, SMSociety '16. ACM, New York, NY, USA, pp. 6:1–6:10.

25. **Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., & Ananiadou, S.** (**2013**). Distributional semantics resources for biomedical text processing. *Proceedings of Languages in Biology and Medicine*.

26. **Ramage, D., Manning, C. D., & Dumais, S.** (**2011**). Partially labeled topic models for interpretable text mining. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 457–465.

27. **Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., & Inches, G.** (**2013**). Overview of the author profiling task at pan 2013. *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, CELCT, pp. 352–365.

28. **Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W.** (**2015**). Overview of the 3rd author profiling task at PAN 2015. *CLEF*.

29. **Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daeleman, W., et al.** (**2014**). Overview of the 2nd author profiling task at PAN 2014. *CEUR Workshop Proceedings*, Vol. 1180, CEUR Workshop Proceedings, pp. 898–927.

30. **Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B.** (**2016**). Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. *Working Notes Papers of the CLEF*.

31. **Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M.** (**2010**). Classifying latent user attributes in twitter. *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ACM, pp. 37–44.

32. **Rastegar-Mojarad, M., Liu, H., & Nambisan, P.** (**2016**). Using social media data to identify potential candidates for drug repurposing: A feasibility study. *JMIR Res Protoc*, Vol. 5, No. 2. doi:10.2196/resprot.5621.

33. **Rong, X.** (**2014**). Word2vec parameter learning explained. *CoRR*, abs/1411.2738.

34. **Sarker, A. & Gonzalez, G.** (**2015**). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, Vol. 53, pp. 196–207.

35. **Sarker, A., Nikfarjam, A., & Gonzalez, G.** (**2016**). Social media mining shared task workshop. *Proc. Pacific Symposium on Biocomputing*, pp. 581–592.

36. **Segura-Bedmar, I., Martınez, P., Revert, R., & Moreno-Schneider, J.** (**2015**). Exploring

Spanish health social media for detecting drug effects. *BMC Medical Informatics and Decision Making*, Vol. 15, No. 2, pp. 1–9. doi:10.1186/1472-6947-15-S2-S6.

37. **Shaywitz, D. & Mammen, M.** (**2011**). The next killer app. *The Boston Globe*.
38. **Stamatatos, E.** (**2009**). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, Vol. 60, No. 3, pp. 538–556.
39. **Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B.** (**2015**). *Overview of the author identification task at PAN*.
40. **Tutubalina, E. & Nikolenko, S. I.** (**2016**). Automated prediction of demographic information from medical user reviews. *Proc. 4th International Conference on Mining Intelligence and Knowledge Exploration*, Lecture Notes in Artificial Intelligence, Springer.
41. **Wilson, T., Wiebe, J., & Hoffmann, P.** (**2009**). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, Vol. 35, No. 3, pp. 399–433.
42. **Yang, C. C., Yang, H., Jiang, L., & Zhang, M.** (**2012**). Social media mining for drug safety signal detection. *Proceedings of the International Workshop on Smart Health and Wellbeing*, SHB '12, ACM, New York, USA, pp. 33–40. doi:10.1145/2389707.2389714.
43. **Zhang, X., Zhao, J., & LeCun, Y.** (**2015**). Character-level convolutional networks for text classification. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, MIT Press, Cambridge, USA, pp. 649–657.
44. **Zhang, Z., Nie, J.-Y., & Zhang, X.** (**2016**). An ensemble method for binary classification of adverse drug reactions from social media. *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*.
45. **Zheng, R., Qin, Y., Huang, Z., & Chen, H.** (**2003**). Authorship analysis in cybercrime investigation. *Intelligence and Security Informatics*, Springer, pp. 59–73.

*Article received on 14/11/2016; accepted on 17/03/2017.*
*Corresponding author is Elena Tutubalina.*