RESEARCH

Neural Side Effect Discovery from User Credibility and Experience-Assessed Online Health Discussions

Van-Hoang Nguyen*, Kazunari Sugiyama, Min-Yen Kan and Kishaloy Halder

*Correspondence:
vhnguyen@u.nus.edu
School of Computing, National University of Singapore, 13 Computing Drive, 117417, Singapore
Full list of author information is available at the end of the article

Abstract

Background: Health 2.0 allows patients and caregivers to conveniently seek medical information and advice via e-portals and online discussion forums, especially regarding potential drug side effects. Although online health communities are helpful platforms for obtaining non-professional opinions, they pose risks in communicating unreliable and insufficient information in terms of quality and quantity. Existing methods in extracting user-reported adverse drug reactions (ADRs) in online health forums are not only insufficiently accurate as they disregard user credibility and global drug experience, but are also expensive as they rely on supervised ground truth annotation of individual statement. We propose a NEural ArchiTecture for Drug side effect prediction (NEAT), which is optimized on the task of drug side effect discovery based on a complete discussion while being attentive to user credibility and experience, thus, addressing the mentioned shortcomings. We train our neural model in a self-supervised fashion using ground truth drug side effects from `mayoclinic.org`. NEAT learns to assign each user a score that is descriptive of their credibility and highlights the critical textual segments of their post.

Results: Experiments show that NEAT improves drug side effect discovery from online health discussion by 3.04% from user-credibility agnostic baselines, and by 9.94% from non-neural baselines in term of F_1. Additionally, the latent credibility scores learned by the model correlate well with trustworthiness signals, such as number of "thanks" received by other forum members, and improve credibility heuristics such as number of posts by 0.113 in term of Spearman's rank correlation coefficient. Experience-based self-supervised attention highlights critical phrases such as mentioned side effects, and enhances fully supervised ADR extraction models based on sequence labelling by 5.502% in term of Precision.

Conclusions: NEAT considers both user and globally reported experiences in online health forums, making feasible a self-supervised approach to side effect prediction for mentioned drugs. The derived user credibility and attention mechanism are transferable and improve downstream ADR extraction models. Our approach enhances automatic drug side effect discovery and fosters research in several domains such as clinical studies.

Keywords: online health communities; drug side effect discovery; credibility analysis; deep learning; natural language processing

1 Background

Seeking medical opinions from online health communities has become popular: 71% of adults aged 18–29 (equivalent to 59% of all U.S. adults) reported consulting online health websites for opinions [1]. These opinions come from an estimated twenty to one hundred thousand health-related websites [2], inclusive of online health communities that network patients with each other to provide information and social support [3]. Platforms such as HealthBoards[1] and MedHelp[2] feature users reporting their own health experiences, inclusive of their self-reviewed drugs and medical treatments. Hence, they are valuable sources for researchers [4, 5].

Although patients use these platforms to access valuable information about drug reactions, there are challenges to their effective, large-scale use. There is lexical variation where users describe the same side effect differently. For example, *dizziness* can be expressed as *giddiness* or *my head is spinning*, posing difficulty to most feature-based or keyword matching approaches. Separately there are valid concerns regarding credibility of user-generated contents to be harvested at large which research has shown to be of variable quality and should be approached with caution [6, 7, 8, 9]. One proxy indicator for information quality is the author's trustworthiness [10]. In the context of social media or online forums, user trustworthiness is often approximated via ratings from other users, *i.e.,* number of thanks or upvotes [11], or via their consistency of reporting credible information [12, 13]. In addition to credibility, forum members also offer expertise thanks to their own experience – with prescriptions in particular – and facilitate responses to drug queries [14]. For instance, while reporting expected side effects for a specific treatment, patients with long-term use of certain drugs can be a complementary source of information:

While my experience of 10 years is with Paxil, I expect that Zoloft will be the same. You should definitely feel better within 2 weeks. One way I found to make it easier to sleep was to get lots of exercize [sic]. Walk or run or whatever to burn off that anxiety. – User 3690.

The above is an answer to a thread asking for expected side effects for depression treatment with `Zoloft`. User 3690's history of active discussion on other anti-depressants such as `Lexapro` and `Xanax` lends credibility to them being an authority on depression treatments. We noticed that `Zoloft` (mentioned in the thread) shares many common side effects with the other two anti-depressants: "*changed behavior*," "*dry mouth*," and "*sleepiness or unusual drowsiness*." as illustrated in Table 1. Many such examples suggest that drugs which are often prescribed together for the same treatment, such as anti-depressants, are likely to be discussed within a same thread and share common side effects. In addition, users who have experienced certain drug reactions are more outspoken and active on those discussions involving drugs of similar side effects. These signals arise from the rich context of online health information, hence, we expect systems to explore beyond individual statements. Specifically, they should consider the complete discussion content as well as the global experience of each involved users, in order to discover drug side effects or extract adverse drug reactions (ADRs).

We argue that modeling user expertise from previously experienced side effects is more robust compared to general user profile and engagement features [13, 14] as their expertise provides more meaningful signals for side effect discovery.

[1] `https://www.healthboards.com/`
[2] `https://medhelp.ord/`

To the best of our knowledge, there is no previous work that incorporates user expertise in side effect discovery in discussion forums at either the thread or post level. In this work, given discussions in online health communities, we propose a novel end-to-end neural architecture that jointly models each author's credibility, their global experience and their post's textual content to discover the side effect of unseen drugs. We optimize the model on a self-supervised task of predicting for side effect of mentioned drugs, where ground truth is accessible. Our key observation is that users can be grouped into clusters that share the same expertise or interest in certain drugs, possibly due to their common treatment or medical history. We incorporate this critical observation into our user model in representing a post's content via a cluster-sensitive attention mechanism [15]. We also follow general definition of truth discovery and let the model learn a credibility score that is unique to every user and descriptive of their trustworthiness. Our experiments include an overall ablation study to validate the significance of the improvements achieved by each model component. Our proposed approach extends our former work [16] by conducting a correlation study that analyzes the representativeness of credibility scores obtained by self-supervised learning as well as a comparison of our attention-based side effect mention extraction and sequence labeling approaches.

We summarize our contributions as follows:

- We propose a NEural ArchiTecture, NEAT, that captures user expertise and credibility, the semantic content of individual posts and the holistic thread to improve side effect discovery from online health discussions. Its major principle that assesses user credibility and experience can be easily generalized across different neural attentional encoders.
- We formulate a self-supervised task of side effect prediction of mentioned drugs for the proposed network to jointly optimize its components.
- We conduct experiments to verify the validity of our learned credibility and the robustness of self-supervised attention-based extraction, comparing against fully supervised sequence labeling baselines.

2 Related Work

We first review existing approaches to drug side effect discovery from health forums and social media. Next, we examine how these works incorporate user credibility and expertise in their learning objective. Finally, we justify our choice of neural architecture by discussing its modeling capability of context-rich structures such as online discussion.

Drug Side Effect Discovery. Existing methods for drug discovery from online content extract drugs at post and statement level. ADR mining systems typically include a named entity recognition (NER) model and a relationship or semantic role labeling model [17, 18]. Recent neural approaches address lexical variation in user-generated content – the difficulty faced by traditional keyword matching and rule-based approaches – to improve recognition and labeling components [19, 20]. Distributed word representations [21] [22] constructed from context can capture semantics based on the hypothesis that synonyms often share similar contextual words. For example, "headache" and "cephalea" will have close representations if

they share many contextual words like "head" or "pain". These representations are naturally integrated into neural sequential models [23] [24]. However, supervised sequence labeling or mention extraction approaches require laborious annotations at the word (token) level, and are only capable of discovering side effects that are explicitly present in the text. Expert supervision or additional semantic matching models are also required to map such recognized text segments to standardized vocabulary or thesaurus [25]. In contrast, our proposed self-supervised task formulation discovers the aggregated side effects of mentioned drugs for each community discussion by considering the whole thread's content. The list of discussed drugs are tagged by forum moderators or obtained by pattern matching. This learning design not only effectively alleviates the need for expensive, finer-grained annotations but also allows for the prediction of side effects not explicitly mentioned in the discussion.

User Credibility and Expertise Integration. Credibility is of the utmost concern in large scale knowledge harvesting [8, 26, 27]. Previous works on side effect discovery from individual statement or post derive information credibility by verifying a statement's mentioned side effects against ground truth drug-side effect databases, and user credibility by measuring the percentage of their credible statements [13, 28]. Our approach to side effect discovery from discussion content by jointly modeling multiple posts and their authors does not derive statement credibility and derives user credibility differently. We assign each user a positive score that is used to weight their post content in representing the discussion's holistic content. Such weighted summation is mathematically shown in the Appendix A to conform to the general principle of truth discovery, where sources providing credible information should be assigned higher credibility scores, and the information that is supported by credible sources will be regarded as true [10]. Although our dataset does not provide any ground truth for user trustworthiness, we followed the previous usage of ratings or upvotes in online forums and adopted number of "thanks" received from other forum members [11] as our proxy for user trustworthiness. Previous works have modeled user expertise based on user profiles such as demographics; activity features such as posting frequency and posting pattern through time series and network analysis [13, 14]. As shown in an earlier example in section 1, modeling user expertise from previously experienced side effects better captures author authoritativeness for certain side effects and is universally applicable to any online platform.

Modeling Online Discussion Content and Structure As our work makes use of the rich topographical properties of online communities, we briefly review approaches for modeling textual content and post-thread discussion structure. Previous works use probabilistic graphical models implicitly to represent textual content (especially, topic modeling) as bag-of-words [25] [29] or stylistic or linguistic features [13]. Such lightweight representation are well-suited in moderately short contexts, *i.e.,* sentences or posts. However, in terms of modeling long discussions consisting of multiple posts, state-of-the-art models for Community Question Answering feature hierarchical neural architectures [30, 31, 32]. In term of encoding

text, sequential encoders such as Long Short-Term Memory (LSTM) [33] or Convolutional Neural Networks (CNN) [23] are capable of encoding long-term dependencies and semantic expressiveness by leveraging word embeddings. In terms of encoding hierarchical structures such as community discussions consisting of post- and thread-level features, neural architectures allow for straightforward and efficient integration of multiple learning objectives. In addition, our neural architecture, NEAT, incorporates attention mechanism that focus on essential phrases while encoding post content, and joint user credibility learning while optimizing for the side effect discovery objective.

3 Methods

Basic Terminology. To ensure a consistent representation, we define some terms and formalize them as follows:

- A *drug* d has a set of side effects, $S_d = \{s_1, s_2, \ldots\}$
- A *post* p is a message in online forums and contains a sequence of words. Post p is written by a *user* u and belongs to a *thread* t.
- A *user* u is a member of an online community. He/She participates in a list of threads, *i.e.,* $T_u = \{t_1, t_2, \ldots, t_l\}$ by writing at least one post in each thread. We use the terms *user* and *author*, as well as *user experience* and *user expertise* interchangeably. Each user is characterized by their credibility and expertise. Credibility w_u of user u reflects the probability of user u provide trustworthy or helpful information, and is approximated from the number of "thanks" given from other forum members.
- A *thread* t (see Table 2) is an ordered collection of post–user pairs, $Q_t = [(p_1, u_1), (p_2, u_2), \ldots, (p_n, u_n)]$.
  Every thread discusses the treatment for a particular condition and entails a list of prescribed drugs $D_t = \{d_1, d_2, \ldots, d_m\}$. Hence, every thread has a list of aggregated potential side effects defined as $S_t = S_{d_1} \cup S_{d_2} \cdots \cup S_{d_m}$.

Task Definition. *Drug side effect discovery from discussions* is the task of assigning the most relevant subset of potential side effects to threads discussing certain drugs, from a large collection of side effects. We view the drug side effect discovery problem as a multi-label classification task. In our setting, an instance of item–label is a tuple $(\boldsymbol{x}_t, \boldsymbol{y})$ where $\boldsymbol{x}_t$ is the feature vector of thread t derived from its list of post–user pairs Q_t and $\boldsymbol{y}$ is the side effect label vector *i.e.*, $\boldsymbol{y} \in \{0,1\}^S$, where S is the number of possible side effect labels. Given training instances, we train our classifier to predict the list of drug side effects in unseen threads discussing unseen drugs.

Formal Hypothesis. Given a thread t with Q_t, we hypothesize that considering the credibility and experience of user $u \in (p, u) \in Q_t$ improves the quality of feature representation in thread t, resulting in better drug side effect discovery performance.

Self-supervised Drug Side Effect Discovery. We propose a self-supervised learning objective. Instead of relying on the identical and independently distributed

assumption of fully supervised learning, we construct the dataset from threads that can discuss a set of common drugs. We look up the side effects of these mentioned drugs via a drug – side effect medical database obtained from Mayo Clinic portal. Our self-supervised task explores discussion-based side effect discovery and alleviates the need for finer-grain annotation compared with existing approach of statement-based side effect discovery. We also propose our neural architecture, NEAT, that jointly models user credibility, expertise and text content with attention while optimizing for the self-supervised objective. The network has three major components: user expertise representation with rich multi-dimensional vectors; cluster-sensitive attention being capable of focusing on relevant phases for post content encoding improvement; and credibility weighting mechanism which effectively learns to assign credibility score to each user based on their content and enhances thread encoding. Their implementation will be discussed in the following sections. Figure 1 shows the detailed network architecture of our model.

User Expertise Representation (UE). We embed each user $u \in U$ as a vector $\boldsymbol{v}_u$ so that the vector captures user u's experience with certain side effects. As each user u participates in the threads T_u, entailing a list of experienced side effects, we derive user side effect experience vector $\boldsymbol{v^*}_u \in \mathbb{R}^{|S|}$ where S is the set of all possible side effects and $v^*_{u_i} = n_{u_i}$ where user u has discussed i^{th} side effect in n_{u_i} threads. We obtain a user drug experience matrix $\boldsymbol{M}^* \in \mathbb{R}^{|U|\times|S|}$ where j^{th} row of $\boldsymbol{M}^*$ denotes user side effect experience vector of j^{th} user. To avoid learning from sparse multi-hot encoded representations and to improve the model's scalability with the number of side effects, we perform dimensionality reduction, specifically principal component analysis (PCA) [34], to our experience matrix $\boldsymbol{M}^*$ obtained from training set. Figure 4 shows percentage of variance explained versus number of included principal components. Since our PCA plots do not show significant improved percentage of variance explained beyond 100 components, we use $g = 100$ components, reducing our original $\boldsymbol{M}^* \in \mathbb{R}^{|U|\times|S|}$ to user expertise matrix $\boldsymbol{M} \in \mathbb{R}^{|U|\times g}$.

User Cluster Attention (CA). We make an assumption via observations that users in online health communities can be effectively grouped into clusters based on their previous side effect experience. The advantages of clustering users is 2-fold. First, since users in the same clusters share certain parameters, they are jointly modeled and more active forum members leverage less active ones. Secondly, clustering efficiently reduces the number of parameters and improves optimization. We apply K-means – a distance-based unsupervised clustering algorithm [35] – to binary user experience vectors $\boldsymbol{v^*_u}$ after normalization. By using cosine similarity, the algorithm effectively groups users with a high number of co-occurred side effects in the same cluster. To determine the number of clusters c, we plot the silhouette scores against the number of clusters and observe the sharp drop after $c = 7$. We observe the average silhouette score of 0.57 for our choice of $c = 7$, indicating that users are moderately matched to their own groups and separated from other groups. The top 5 most common side effects in each clusters are shown in Table 3.

In a larger domain of Natural Language Processing, attention has become an integral part for modeling text sequences [36] [37]. By learning to focus on essential

text segments, attention allows text encoders to capture long term semantic dependencies with regard to auxiliary contextual information [38] [39]. In our related task of ADR mentions extraction, attention has been adopted recently in neural sequence labelling models [19] [40], resulting in promising improvement. Inspired by the concept, we enhance text encoding with user expertise attention. Even though the attention is adjusted to the non-extractive self-supervised task of thread-level drug side effect discovery, we hypothesize that our model learns to highlight the mentioned accurate side effects, and can be used as a self-supervised baseline for side effect extraction. Based on the previously obtained clustering results, we assign a learnable cluster attention vector for each user group and incorporate their expertise into the text encoding process.

Post Content Encoding. The network takes the content of a thread t as input, which is a list of post–user pairs Q_t. Post p_i of pair $(p_i, u_i) \in Q_t$ consists of a sequence of words $x_{p_i} = \{w_1, \ldots, w_n\}$. We seek to represent a post p_i as a vector $\boldsymbol{v}_p$ that effectively captures its semantics through an encoding function $f(x_{p_i})$ modeled by a neural text encoding module (the blue boxes in Figure 1). We embed each word into a low dimensional vector and transform the post into a sequence of word vectors $\{\boldsymbol{v}_{w_1}, \boldsymbol{v}_{w_2}, \ldots, \boldsymbol{v}_{w_n}\}$. Each word vector is initialized using pre-trained GloVe [22], and each out-of-vocabulary word vector is initialized randomly. We make use of modularity – a major advantage of neural architectures and design the post content encoder as a standalone component that can be easily updated with any state-of-the-art text encoder. In this work, we provide two neural text encoders: long-short term memory (LSTM, see Figure 2) [33] and convolutional neural networks (CNN, see Figure 3) [23], both of which incorporates attention mechanism.

A bi-directional LSTM encodes the word vector sequence and outputs two sequences of hidden states: a forward sequence, $H^f = \boldsymbol{h}_1^f, \boldsymbol{h}_2^f, \ldots, \boldsymbol{h}_n^f$ that starts from the beginning of the text; and a backward sequence, $H^b = \boldsymbol{h}_1^b, \boldsymbol{h}_2^b, \ldots, \boldsymbol{h}_n^b$ that starts from the end of the text. For many sequence encoding tasks, knowing both past (left) and future (right) contexts has proven to be effective [41]. The states h_i^f and h_j^b in the forward and backward sequences are computed as follows:

$$\boldsymbol{h}_i^f = LSTM(\boldsymbol{h}_{i-1}^f, \boldsymbol{w}^i),\ \boldsymbol{h}_j^b = LSTM(\boldsymbol{h}_{j+1}^b, \boldsymbol{w}^j),$$

where $\boldsymbol{h_i^f}, \boldsymbol{h_j^b} \in \mathbb{R}^e$, and e are the number of encoder units. We derive the cluster attention vector as $\boldsymbol{v}_{a_i} \in \mathbb{R}^e$ for each user c_i. Given a forward sequence $H^f = \boldsymbol{h}_1^f, \boldsymbol{h}_2^f, \cdots, \boldsymbol{h}_n^f$ and backward sequence $H^b = \boldsymbol{h}_1^b, \boldsymbol{h}_2^b, \cdots, \boldsymbol{h}_n^b$ of hidden post states p written by user u belonging to cluster c_i, the corresponding w_{a_j} weights each hidden state $\boldsymbol{h}_j^f$ and $\boldsymbol{h}_j^b$ of both sequences based on their similarity with the attention vector are:

$$w_{a_j} = \frac{exp(\boldsymbol{v}_{a_i} \boldsymbol{h}_j)}{\sum_{l=1}^{n} exp(\boldsymbol{v}_{a_i} \boldsymbol{h}_j)}. \quad (1)$$

The intuition behind Equation (1), inspired by [36], is that hidden states which are similar to the attention vector $\boldsymbol{v}_{a_i}$ should be paid more attention to; hence are weighted higher during document encoding. $\boldsymbol{v}_{a_i}$ is adjusted during training to

capture hidden states that are significant in forming the final post representation. w_{a_j} is then used to compute forward and backward weighted feature vectors:

$$\boldsymbol{h}^f = \sum_{j}^{n} w_{a_j} \boldsymbol{h}_j^f, \quad \boldsymbol{h}^b = \sum_{j}^{n} w_{a_j} \boldsymbol{h}_j^b. \tag{2}$$

We concatenate the forward and backward vectors to obtain a single vector, following previous bi-directional LSTM practice [42].

Our choice of CNN-based encoder is inspired by existing works [23] [43]. A convolution block k consists of two sub-components: a convolution layer and a cluster attention layer. In the convolution layer, a kernel of window $s, 0 < s < n$ of weight $\boldsymbol{W}$ is used to generate the hidden representation $\boldsymbol{h_j^k}$ for the word embeddings $\boldsymbol{v_{w_{i-s+1}}}, \cdots, \boldsymbol{v_{w_i}}$ as

$$\boldsymbol{h_j^k} = CONV(\boldsymbol{W}, [\boldsymbol{v_{w_{i-s+1}}}, \cdots, \boldsymbol{v_{w_i}}]) \tag{3}$$

where $CONV(\cdot)$ is the convolution operation described in [23]. In the cluster attention layer, we first derive the attention weight w_{a_j} for each hidden representation $\boldsymbol{h_j^k}$ similarly to the LSTM-based encoder. Attention weighted pooling is used to obtain the convolution block output as follows:

$$\boldsymbol{h}^k = \sum_{j}^{n} w_{a_j} \boldsymbol{h}_j^k \tag{4}$$

Since we use multiple convolution blocks of different kernel sizes, the final post representation is the concatenation of K block outputs $\boldsymbol{h}^k$.

Thread Content Encoding with Credibility Weights (CW). For every post–user pair (p_i, u_i) at thread t, we first compute feature vector $\boldsymbol{v}_{p_i}$ for post p_i. It is then concatenated with user u_i's expertise vector $\boldsymbol{v}_{u_i}$ to form post–user complex vector $\boldsymbol{v}_{u_i}^p$. This user-post complex is weighted by a user credibility $e^{w_{u_i}}$, where w_{u_i} initially set to 0 per user and updated while training for the self-supervised side effect discovery objective. We implement credibility learning according to the general intuition from the truth discovery literature: users who give quality posts, on which the model can solely base to make correct predictions, are given a higher credibility. We also exploit this credibility score to encode the thread representation more effectively by placing emphasis on the content of credible users. A representation of thread that meets the above description is the weighted sum of each post–user complex vector:

$$\boldsymbol{v}_t = \sum_{i=1}^{n} \boldsymbol{v}_{u_i}^{p*} = \sum_{i=1}^{n} e^{w_{u_i}} \boldsymbol{v}_{u_i}^{p} \tag{5}$$

Multi-label Prediction: We feed the thread content representation v_t through a fully connected layer whose outputs can be computed as follows:

$$\boldsymbol{s}_t = \boldsymbol{W} tanh(\boldsymbol{v}_t) + \boldsymbol{b}, \tag{6}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are weights and biases of the layer. The output vector $\boldsymbol{s}_t \in \mathbb{R}^{|S|}$ is finally passed through a sigmoid activation function $\sigma(\cdot)$, and trained using cross-entropy loss L defined as follows:

$$\begin{aligned} L = \frac{1}{T}\sum_{t=1}^{T}\{\boldsymbol{y}_t \cdot log(\sigma(\boldsymbol{s}_t)) + (1-\boldsymbol{y}_t)\cdot log(1-\sigma(\boldsymbol{s}_t))\} \\ + \lambda_1\sqrt{\sum_u \boldsymbol{v}_u^2} + \lambda_2 \sum_i |\boldsymbol{w}_{u_i}| \end{aligned} \tag{7}$$

We adopt regularization that penalizes the training loss with the user experience matrix's $L2$ norm by a factor of λ_1 and the user score vector $\boldsymbol{w}_u$'s $L1$ norm by a factor of λ_2. The loss function is differentiable, thus trainable with Adam optimizer [44]. During our gradient-based learning, user u_i's credibility score w_{u_i} can be updated by calculating $\frac{\partial L}{\partial w_{u_i}}$ by back-propagation (see Appendix A).

4 Results

We conduct experiments to validate the effectiveness of our proposed model. More specifically,

1 We design an ablation study to highlight the effectiveness of each component of NEAT in our self-supervised side effect prediction.
2 We verify the representativeness of the learned credibility scores via correlation analysis and ranking metrics using number of "thanks" received by other forum members as the trustworthiness proxy.
3 We compare the model's performance in unseen drug side effect discovery against non-neural baselines.
4 We examine the applicability of cluster attention in side effect mention extraction from user posts both at the macroscopic and microscopic levels.

Dataset and Experiment Settings. We conduct our experiments on the same dataset as [13] including 15,000 users and 2.8 million posts extracted from 620,510 HealthBoards[1] threads. The ground truths for self-supervised learning are defined as side effects of mentioned drugs in the discussion. As annotating such amount of posts is expensive, drug side effects are extracted from Mayo Clinic's Drugs and Supplements portal[3] and are used as surrogates for potential drug reactions. From the original dataset, we only extract threads that are annotated with drugs and their side effects, along with the lists of contained posts and corresponding users. Table 4 shows some statistics of our dataset.

For CNN encoder we adopted the work by Kim (2014) [23] and used 3 kernels of sizes 3, 4, 5 with output channel size = 100. For Bi-LSTM we use 1 layer with hidden state size = 32.

[3] `https://www.mayoclinic.org/drugs-supplements`

We used Natural Language Toolkit [4] for tokenizationa and and stop-word elimination before representation modeling. We perform cross-validation (8, 1, and 1 folds for training, validation, and testing, respectively). We perform PCA and K-means clustering on training set, using scikitlearn's built-in modules [45], 100 principal components ($g = 100$). All models are trained using Pytorch[5] library. We release our codes at [6].

Ablation Study. We include each component in Section 3 to the architecture at a time and verify the incremental enhancement. We implement both CNN and LSTM-based text encoders to confirm the consistent improvement across different neural encoders. The ablated baselines and the full models are as follows:

- **Vanilla**: We implement a neural text encoder baseline without any proposed component.
- **Weighted Post Encoder (WPE)**: We construct thread representation by summing each of its post–user complex vector weighted by user credibility.
- **Weighted Post Encoder with User Expertise (WPEU)**: We concatenate user expertise with post vector to create post–user complex vector.
- **NEAT**: We incorporate all three components, UE, CW and CA, as described.

Table 5 shows the precision, recall, and F_1 obtained by our method and the four baselines. We also report the performance of baselines implemented UE and CA individually in Table 11.

User Credibility Analysis. We discuss how descriptive the credible users assigned by the model are to our common notion of trust-worthy users in online communities. We employ the number of "thanks" received by other community member as the proxy for a user's credibility, at both global, forum-wise scope and local, thread-wise scope. Specifically, at forum-wise scope, we measure Spearman's rank correlation coefficient to examine how our output user scores approximate the ordering of user trustworthiness. At thread-wise scope, we examine how accurate our output user scores are in ranking trustworthy respondents within a single discussion by measuring both Spearman's coefficient and nDCG@2 with regard to the ordering provided by our credibility proxy. We find the forum-wise ranking metrics meaningful as answer ranking, based on user credibility in our case, is a well-formulated task in Community Question Answering [46, 47, 48]. The measurement results are presented in Table 6.

Drug Side Effect Discovery. We test the performance of NEAT in the task of Drug Side Effect Discovery. Specifically, the model has to predict the side effects of one of five unseen drugs based on their discussions as a whole. Such task is necessary to verify that our self-supervised objective of predicting for side effects of discussed drugs generalizes well to drugs that have not been discussed in training data. We highlight the performance of an end-to-end neural architecture against Random Forest (RF) – a non-neural baseline trained on bag-of-word text representations

[4] {https://www.nltk.org/index.html}
[5] https://pytorch.org/
[6] https://github.com/nguyenvanhoang7398/NEAT

of thread content. Additionally, we examine a baseline, uNEAT, where the user identities are randomized in order to verify that NEAT effectively considers both user credibility and expertise. The results of drug side effect discovery on Ibuprofen, Levothyroxine, Metoformin (Table 7), and Omeprazole, Alprazolam (Table 8) are reported.

Side Effect Extraction with Cluster Attention. As discussed in section 3, our employed attention mechanism not only offers a better textual encoding capacity but also locates the informative segments of user-generated content. This concept is well-aligned with ADR mention extraction, which is mainly modeled as a sequence labeling problem. We conduct experiments to examine the effectiveness of CNN-NEAT's Attention in locating the text segments containing the correct side effects. We benchmark our results against a lexicon-based tagging using UMLS thesaurus for medical terms [49] and a state-of-the-art neural side effect extractor [19] which was supervisedly trained to identify side effect mentions in social media contents. In this task, correctly predicting the positive side effects is of the utmost importance, hence we benchmark the text segments extracted from CNN-NEAT's Attention against two mentioned baselines on precision metric. The experiment results on Ibuprofen, Levothyroxine, Metoformin, Omeprazole and Alprazolam are reported in Table 9.

5 Discussion

Ablation Study. Firstly, all of the three models that apply credibility weighting (CW) – WPE, WPEU, and NEAT outperform both and LSTM-Vanilla and CNN-Vanilla baselines. Specifically, in LSTM-Vanilla, solely weighting each post by its author credibility improves the performance of naive post encoder by 2.03%, 2.74% and 1.68% on precision, recall, and F_1, respectively. We observe a similar margin of improvement from CNN-Vanilla. These demonstrate the effectiveness of accounting for author credibility when encoding thread content, improving side effect prediction.

Improvements by incorporating user experience (UE) are also testified in both neural encoders. In LSTM-based models, adding UE (LSTM-WPEU vs. LSTM-WPE) improves recall by 6.57% and 3.93% in F_1. Again, the CNN-based counterpart, CNN-WPEU shows similar performance trends. On a macro scale, these statistics indicate that our model successfully learns to include more side effects in its prediction, where many are relevant to the ground truth. This is consistent with our hypothesis that considering author experience of each post is effective in predicting out-of-context side effects.

Applying cluster-sensitive attention (CA) in combining both LSTM and CNN's hidden states also improves the performance. In LSTM-based systems, we observe that adding CA (LSTM-NEAT vs. LSTM-WPEU) uniformly improves all retrieval metrics, where again, the CNN-based counterpart, CNN-NEAT, demonstrates similar performance improvements. Although CNN-NEAT and its ablated baselines obtain higher performance than the LSTM counterparts, when measuring relative improvement, the gains are comparable. This confirms the consistent improvement of our proposed components across different neural encoders. According to the

macroscopic analysis of results in Table 5, we generally conclude that all of the three components in our proposed architecture, namely, CW, UE, and CA give a positive impact on the overall performance of the model. We observe consistent improvements in F_1 after adding each component, and this solidates our stated hypotheses. The significance of these findings were verified by one-tail t-test of $p < 0.05$.

User Credibility Analysis. Results at both thread and forum level show that user scores assigned by NEAT reasonably approximate our credibility proxy, *i.e.* the number of "thanks" given by other forum users. Regarding ranking users by their helpfulness within a thread, NEAT's credibility improves heuristics such as post or question frequency marginally by 0.0044 in term of nDCG@2 and moderately by 0.0180 in term of Spearman's coefficient. Regarding ranking users by their helpfulness in the whole forum, we report a more significant improvement of 0.1131 in term of Spearman's coefficient from the closest performing baseline of post frequency. These results verify the representativeness of the credibility scores obtained self-supervisedly by our system.

Drug Side Effect Discovery. Both neural methods, namely uNEAT and NEAT, outperform non-neural approach based on Bag-of-word vectors on all metrics and across all drugs. Specifically, NEAT improves RF by 9.94% in F_1 on average across five different drugs. This confirms the modeling capability of our neural network and justifies its application to our task. We also observe a decrease in performance while optimizing NEAT without being aware of users. Such shortcoming is prevalent in all side effect discovery settings, ranging from 0.7% F1 score decrease in Metformin to 4.9% decrease in Ibuprofen. Overall, being aware of user experience and expertise improves drug side effect discovery by 3.04% in F_1 on average across five different drugs. This again confirms our hypothesis that considering the credibility and experience of users improves drug side effect discovery performance in online health communities.

Side Effect Extraction with Cluster Attention. We notice improvement of CNN-NEAT's Attention from both baselines across most drugs with the exception of Omeprazole. At macro level, positive precision scores confirm CNN-NEAT's emphasis on critical text segments, *i.e.* those containing correct drug side effects. The significant improvement in most cases also suggests the selectiveness of CNN-NEAT's Attention. Unlike the two proposed baselines which extract any side effect mentions, CNN-NEAT's Attention selectively emphasize on the correct ADR. We examine this hypothesis at micro level in Table 10. UMLS tagging identifies any phrase having medical nuance, *i.e. pain* and *back*, and potentially forms side effects that are not actually mentioned, *i.e. back pain*, whereas both neural methods, Neural Extractor and NEAT, that model textual semantics are able to dismiss such trivial mentions. *Discomfort*, although correct, is questionably an intentionally reported side effects and is also dismissed by both Neural Extractor and CNN-NEAT. We attempt to explain CNN-NEAT's decision to dismiss *restlessness* as the attention was derived from each cluster's experienced side effects, in which there is the *weak*

side effects which are semantically contradicting to *restlessness*. Although having not fully covered all side effects, all mentions extracted by CNN-NEAT are accurate, giving it the highest precision amongst the three considered models. Specifically, CNN-NEAT's Attention improves neural ADR extraction model [19] by 5.502% and UMLS tagging by 3.974% on average in term of Precision across five different drugs. Despite being derived from a self-supervised objective, CNN-NEAT's Attention offers helpful indications for attention-based models and a strong baseline for ADR extraction.

Limitations. We foresee a limitation arisen from our design of user's credibility defined by Equation (5) as well as our choice of credibility proxy using the number of "thanks". A user's credibility can be damaged if their posts do not directly help with predicting the correct side effects. This assumption is questionable when users are asking for some information instead of giving answers without any intend to give misleading information. In contrast, we also observe the cases where users receive thanks for giving helpful information such as suggesting nutritious diet or healthy lifestyle without mentioning any relevant side effects. We recognize the limitation of our model where users without malicious intent are possibly assigned a low credibility score. This case of "failure" can explain why some users are assigned low to moderate credibility despite their high number of "thanks". However, our definition makes sure that the credibility learning mechanism does not express the opposite adverse behavior of assigning high credibility to untrustworthy users.

We realize the limitation arisen from annotating a post's side effects solely from looking up the mentioned drugs in the Mayo Clinic database. To generalize the usability and ensure the effectiveness of the learning framework to a broader community such as Medical informatics, we suggest to have these annotations cross-checked with healthcare professionals or pharmacovigilance experts, in order to ensure the correlation between Mayo Clinic's annotations and the post's actually described side effects.

Overall, our analysis suggests that user credibility scores, although learned in a self-supervised manner, can capture the expected notion of credibility and are descriptive of trustworthiness. Every component of our architecture is also shown to be vital in achieving the highest performance.

6 Conclusion

We have addressed the importance of user experience and credibility in modeling thread contents of online communities, specifically through the task of drug side effect discovery. Our proposed neural architecture, NEAT, suggests a subset of side effects relevant to the mentioned treatment in the given discussion, taking into account the each post content and its author side effect experience via attention mechanism to represent forum discussion better. Mainstream models for drug discovery in online communities have not captured thread content and user experience holistically in an end-to-end optimizable system.

We modeled users' expertise by examining their experience with different side effects, and then grouped the users with similar experience into clusters that share a common attention representation. We also proposed an self-supervised method

which assigns credibility scores to users based on the correctness of their contents and overall improves thread representations. Correlation analysis testifies the representativeness of the scores of the common notion of trustworthiness approximated by number of "thanks" received by other online community members. In addition, our integrated attention mechanism not only enhances textual encoding but also highlights essential text segments and benefits ADR extracting approaches.

We believe that our model is applicable to other domains. We plan to generalize its application to mainstream community question answering or expertise-based thread recommendation for health forum members.

Appendix

A User Credibility Weighting and the general principle of truth discovery

In order to demonstrate the correlation between learned user credibility scores and the general notion of trustworthiness, we derive how user credibility scores are updated after each turn of back-propagation via stochastic gradient descent. The overall loss function in Equation (7) can be rewritten as logistic loss without regularization on a single training example and a single label s as follows:

$$L = log(1 + exp(-y_s(\boldsymbol{w}_s^T tanh(\boldsymbol{v_t}) + b_s))), \tag{8}$$

where y_s is the binary truth for label s, b_s is a classification bias, and $\boldsymbol{w_s} \in \mathbb{R}^{g \times 1}$ is a row of $\boldsymbol{W}$ in Equation (6). $\boldsymbol{w}_s$ is the classification weights of a single label s.

In back-propagation, we update the score w_{u_i} of user u_i based on the gradient calculated by taking the derivative of the loss L with regard to w_{u_i}:

$$\frac{\partial L}{\partial w_{u_i}} = \frac{(1 - tanh^2(\boldsymbol{v_t}))y_s \boldsymbol{w}_s^T \boldsymbol{v}_{u_i}^p e^{w_{u_i}}}{1 + exp(y_s(\boldsymbol{w_s}^T tanh(\boldsymbol{v_t}) + b_s))} = \nabla_{w_{u_i}} L. \tag{9}$$

User score w_{u_i} is updated as follows:

$$w_{u_i}^{t+1} = w_{u_i}^t - \eta \nabla_{w_{u_i}^t} L, \tag{10}$$

where η is the learning rate.

When the prediction is correct, y_s and $(\boldsymbol{w}_s^T tanh(\boldsymbol{v_t}) + b_s)$ share the same sign and $y_s(\boldsymbol{w}_s^T tanh(\boldsymbol{v_t}) + b_s))$ is highly positive, making the denominator highly positive and the overall gradient small. The user score w_{u_i} is minimally updated.

When the prediction is incorrect, y_s and $(\boldsymbol{w_s}^T tanh(\boldsymbol{v_t}) + b_s)$ have different signs and the denominator approaches 1. In the nominator, $\boldsymbol{w}_s^T \boldsymbol{v}_{\boldsymbol{u_i}}^p$ is the prediction if we solely consider the post vector $\boldsymbol{v}_{u_i}^p$ of user u_i.

- If this prediction is correct, which fits our definition of credible user, $y_s \boldsymbol{w}_s^T \boldsymbol{v}_{u_i}^p$ is positive, making the overall gradient positive. User score w_{u_i} is updated in the positive direction and the credibility score $e^{w_{u_i}}$ used to weight user u_i's content increases.
- On the other hand, when the prediction from solely considering the post vector $\boldsymbol{v}_{u_i}^p$ of user u_i is incorrect, indicating a not credible user, $y_s \boldsymbol{w}_s^T \boldsymbol{v}_{u_i}^p$ is negative, and the overall gradient is negative. w_{u_i} is updated in the negative direction and the credibility score $e^{w_{u_i}}$ used to weight user u_i's content decreases.
- The magnitude of the gradient is proportional to $e^{w_{u_i}}$. This indicates that users who are currently learned as credible are most affected by back-propagation when the model's prediction is incorrect.

B Algorithm performance for individual integration of UE and CA.

Table 11 reports the performance of baselines implemented UE, and CA individually.

Abbreviations

Avg.: Average
UE: User Expertise Representation
CA: Cluster-sensitive Attention
CW: Thread Content Encoding with Credibility Weights
CNN: Convolutional Neural Network
LSTM: Long-short Term Memory
F_1 **score**: The harmonic average of the precision and recall
PCA: Principal Component Analysis
#: Number of
NEAT: The neural architecture proposed in this work
uNEAT: The ablated version of NEAT where the user identities are randomized
ADR: Averse Drug Reaction
RF: Random Forest
nDCG@2: normalized Discounted Cumulative Gain at 2

Declarations

Acknowledgements
None declared.

Funding
None declared.

Authors' contributions
All authors conceived of and planned the reported work. KS proposed the problem statement and suggested literature for review. VHN mainly designed the model architecture with the help from MYK and KH. VHN implemented the experiments, and interpreted the results. VHN, KS, MYK took the lead in writing the manuscript with support from KH. All authors discussed the results and commented on the manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
All authors declare that they have no competing interests.

Availability of data and materials
The dataset analysed during the current study was first published in [13] and is publicly available at https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/peopleondrugs/
The source codes of the implementation of this study is publicly available at https://github.com/nguyenvanhoang7398/NEAT
The source codes used to generate the results of Side Effect Extraction in Section 4 is available at https://github.com/nguyenvanhoang7398/NEAT/blob/master/adr_extraction.py

References

1. Fox, S., Duggan, M.: Health Online 2013. Health **2013**, 1–55 (2013)
2. Diaz, J.A., Griffith, R.A., Ng, J.J., Reinert, S.E., Friedmann, P.D., Moulton, A.W.: Patients' Use of the Internet for Medical Information. Journal of General Internal Medicine **17**(3), 180–185 (2002)
3. Johnston, A.C., Worrell, J.L., Di Gangi, P.M., Wasko, M.: Online Health Communities: an Assessment of the Influence of Participation on Patient Empowerment Outcomes. Information Technology & People **26**(2), 213–235 (2013)
4. Leyens, L., Reumann, M., Malats, N., Brand, A.: Use of Big Data for Drug Development and for Public and Personal Health and Care. Genetic Epidemiology **41**(1), 51–60 (2017)
5. Martin-Sanchez, F., Verspoor, K.: Big data in medicine is driving big changes. Yearbook of Medical Informatics **9**(1), 14–20 (2014)
6. Impicciatore, P., Pandolfini, C., Casella, N., Bonati, M.: Reliability of Health Information for the Public on the World Wide Web: Systematic Survey of Advice on Managing Fever in Children at Home. BMJ **314**(7098), 1875 (1997)
7. Peterson, G., Aslani, P., Williams, K.A.: How do consumers search for and appraise information on medicines on the internet? a qualitative study using focus groups. Journal of medical Internet research **5**(4), 33 (2003)

8. Hajli, M.N., Sims, J., Featherman, M., Love, P.E.: Credibility of information in online communities. Journal of Strategic Marketing **23**(3), 238–253 (2015)
9. Poddar, L., Hsu, W., Lee, M.L.: Predicting user reported symptoms using a gated neural network. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (2019). IEEE
10. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J.: A survey on truth discovery. ACM SIGKDD Explorations Newsletter **17**(2), 1–16 (2016)
11. Rains, S.A., Karmikel, C.D.: Health information-seeking and perceptions of website credibility: Examining web-use orientation, message characteristics, and structural features of websites. Computers in Human Behavior **25**(2), 544–553 (2009)
12. Hoang, T., Liu, J., Pratt, N., Zheng, V.W., Chang, K.C., Roughead, E., Li, J.: Authenticity and credibility aware detection of adverse drug events from social media. International journal of medical informatics **120**, 157–171 (2018)
13. Mukherjee, S., Weikum, G., Danescu-Niculescu-Mizil, C.: People on Drugs: Credibility of User Statements in Health Communities. In: Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14), pp. 65–74 (2014)
14. Vydiswaran, V.V., Reddy, M.: Identifying peer experts in online health forums. BMC medical informatics and decision making **19**(3), 68 (2019)
15. Halder, K., Poddar, L., Kan, M.-Y.: Cold Start Thread Recommendation as Extreme Multi-label Classification. In: Proc. of the Workshop on Extreme Multilabel Classification for Social Media Co-located with the Web Conference (WWW'18 Companion), pp. 1911–1918 (2018)
16. Nguyen, V.H., Sugiyama, K., Kan, M.-Y., Halder, K.: Treatment side effect prediction from online user-generated content. In: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, pp. 12–21. Association for Computational Linguistics, Brussels, Belgium (2018). doi:10.18653/v1/W18-5602. https://www.aclweb.org/anthology/W18-5602
17. Sampathkumar, H., Chen, X.-W., Luo, B.: Mining Adverse Drug Reactions from Online Healthcare Forums using Hidden Markov Model. BMC Medical Informatics and Decision Making **14**(1), 91–108 (2014)
18. Liu, Y., Shi, J., Chen, Y.: Patient-centered and experience-aware mining for effective adverse drug reaction discovery in online health forums. Journal of the Association for Information Science and Technology **69**(2), 215–228 (2018)
19. Ding, P., Zhou, X., Zhang, X., Wang, J., Lei, Z.: An attentive neural sequence labeling model for adverse drug reactions mentions extraction. IEEE Access **6**, 73305–73315 (2018)
20. Wunnava, S., Qin, X., Kakar, T., Sen, C., Rundensteiner, E.A., Kong, X.: Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. Drug safety **42**(1), 113–122 (2019)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: In Proc. of the Advances in Neural Information Processing Systems (NIPS 2013), pp. 3111–3119 (2013)
22. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
23. Kim, Y.: Convolutional Neural Networks for Sentence Classification. In: Proc. of the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014), pp. 1746–1751 (2014)
24. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
25. Yates, A., Goharian, N., Frieder, O.: Extracting adverse drug reactions from social media. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
26. Mukherjee, S., Weikum, G.: Leveraging joint interactions for credibility analysis in news communities. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 353–362 (2015). ACM
27. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2173–2178 (2016). ACM
28. Li, Y., Du, N., Liu, C., Xie, Y., Fan, W., Li, Q., Gao, J., Sun, H.: Reliable Medical Diagnosis from Crowdsourcing: Discover Trustworthy Answers from Non-Experts. In: Proc. of the 10th ACM International Conference on Web Search and Data Mining (WSDM 2017), pp. 253–261 (2017)
29. Wang, S., Li, Y., Ferguson, D., Zhai, C.: Sideeffectptm: An unsupervised topic model to mine adverse drug reactions from health forums. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 321–330 (2014). ACM
30. Qiu, X., Huang, X.: Convolutional neural tensor network architecture for community-based question answering. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
31. Zhou, X., Hu, B., Chen, Q., Wang, X.: Recurrent convolutional neural network for answer selection in community question answering. Neurocomputing **274**, 8–18 (2018)
32. Zhang, X., Li, S., Sha, L., Wang, H.: Attentive interactive neural networks for answer selection in community question answering. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
33. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997)
34. Jolliffe, I.T.: Principal Component Analysis and Factor Analysis. Statistical Methods in Medical Research **1**(1), 115–128 (1986)
35. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proc. of the 5th Berkeley Symposium on Mathmatical Statistics and Probability, pp. 281–297 (1967)
36. Luong, M.-T., Pham, H., Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation. In: Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 1412–1421 (2015)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.:

Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
38. Chen, H., Sun, M., Tu, C., Lin, Y., Liu, Z.: Neural sentiment classification with user and product attention. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1650–1659 (2016)
39. Feng, S., Wang, Y., Liu, L., Wang, D., Yu, G.: Attention based hierarchical lstm network for context-aware microblog sentiment classification. World Wide Web **22**(1), 59–81 (2019)
40. Ramamoorthy, S., Murugan, S.: An attentive sequence model for adverse drug event extraction from biomedical text. arXiv preprint arXiv:1801.00625 (2018)
41. Dyer, C., Ballesteros, M., Ling, W., Matthews, A., Smith, N.A.: Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015), pp. 334–343 (2015)
42. Ma, X., Hovy, E.: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pp. 1064–1074 (2016)
43. Yin, W., Schütze, H., Xiang, B., Zhou, B.: Abcnn: Attention-based convolutional neural network for modeling sentence pairs. Transactions of the Association for Computational Linguistics **4**, 259–272 (2016)
44. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proc. of the 3rd International Conference for Learning Representations (ICLR2015) (2015)
45. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (JMLR) **12**(2011), 2825–2830 (2011)
46. Nie, L., Wei, X., Zhang, D., Wang, X., Gao, Z., Yang, Y.: Data-driven answer selection in community qa systems. IEEE transactions on knowledge and data engineering **29**(6), 1186–1198 (2017)
47. Surdeanu, M., Ciaramita, M., Zaragoza, H.: Learning to rank answers on large online qa collections. In: Proceedings of ACL-08: HLT, pp. 719–727 (2008)
48. Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Glass, J., Randeree, B., *et al.*: Semeval-2016 task 3: Community question answering. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 525–545 (2016)
49. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), 267–270 (2004)

Figures

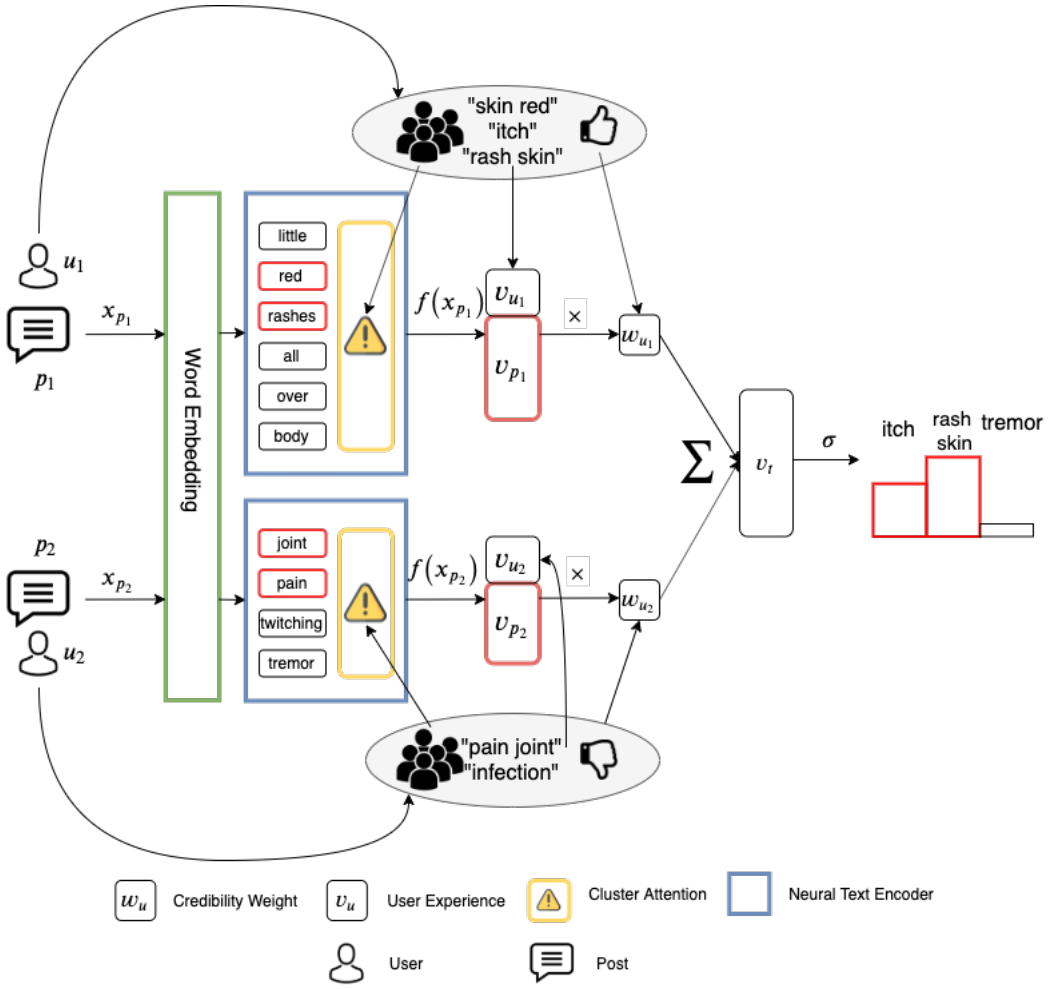

Figure 1: The neural architecture of our proposed NEAT. The "w_u" and "v_u" boxes denote Credibility Weight (CW) component and User Expertise (UE) component. The yellow boxes and blue boxes denote Cluster Attention (CA) component and neural text encoders with attention. The highlighted words in red denoted the text segments that are being attended by the encoder. The "$\times$", "$\sum$", and "σ" symbols denote the multiplication, summation, and sigmoid, respectively.

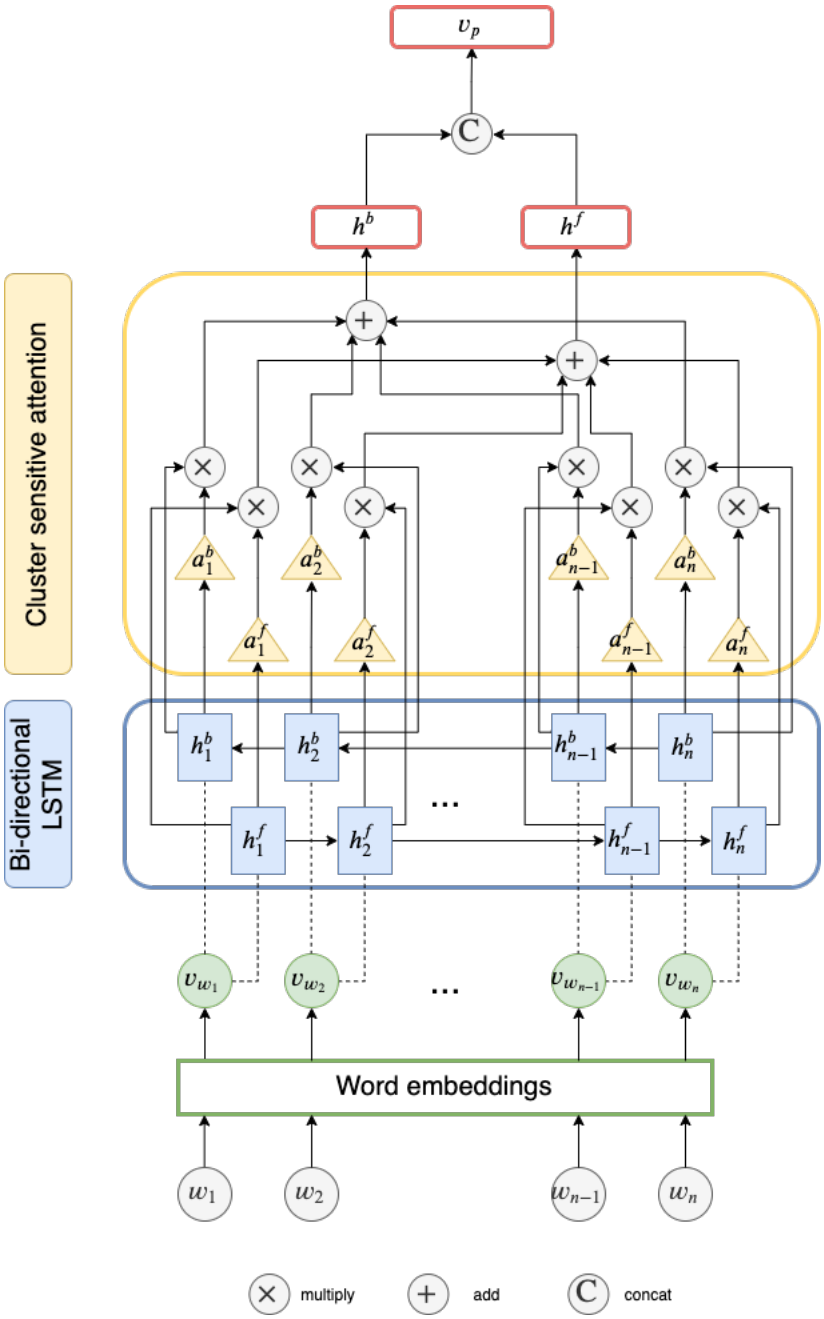

Figure 2: LSTM-based encoder with cluster attention.
The "×" and "+" cells denote the attention-weighted summation described in equation 2. The "C" cell denotes the concatenation of the forward, $\boldsymbol{h^f}$, and backward, $\boldsymbol{h^b}$, hidden states.

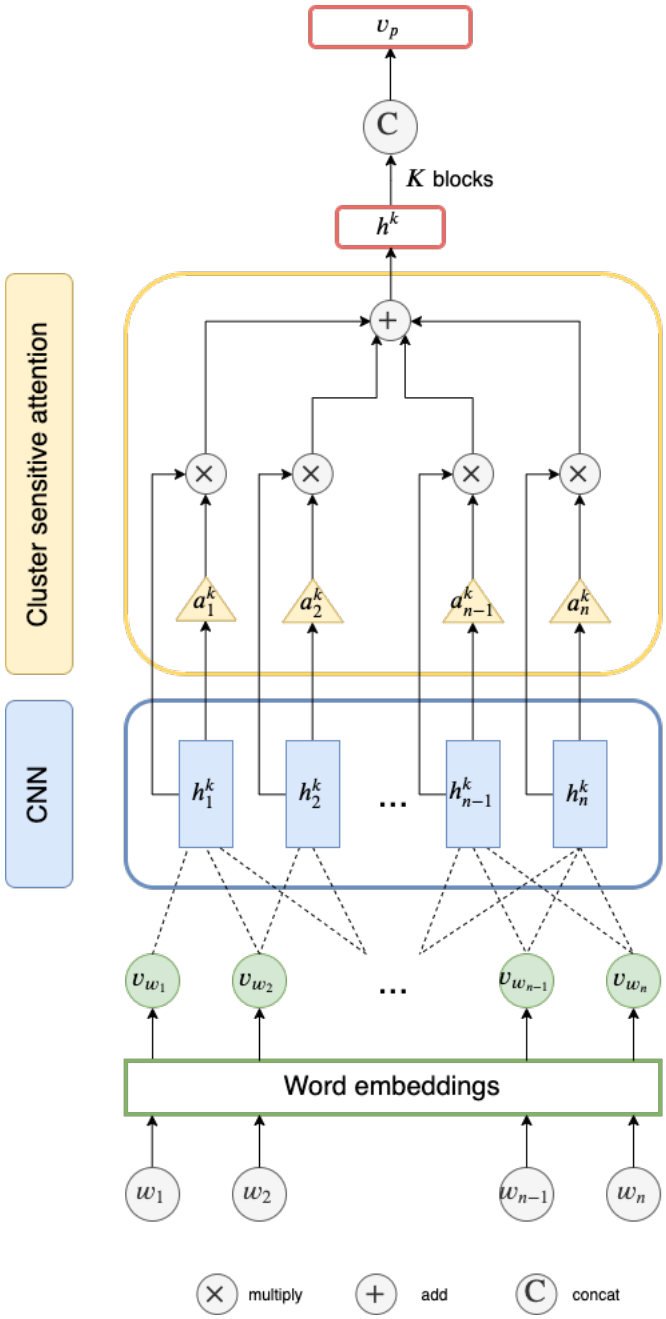

Figure 3: CNN-based Encoder with Cluster Attention.
The "×" and "+" cells denote the attention-weighted summation described in equation 2. The "C" cell denotes the concatenation of the final hidden states of K convolution blocks.

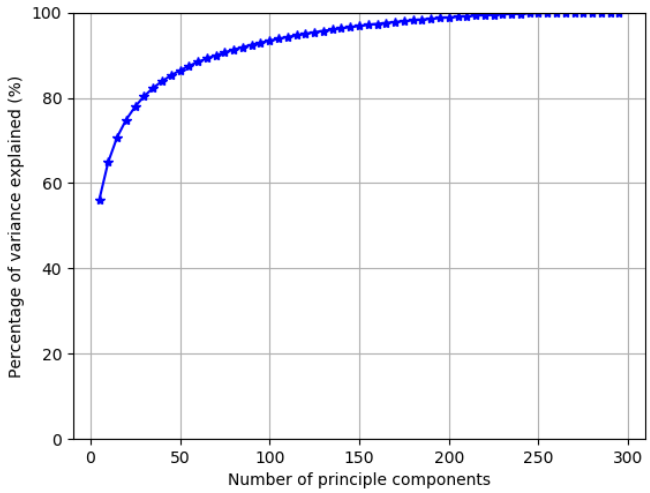

Figure 4: Principal component analysis.
The horizontal axis denotes the number of principle components chosen for PCA, while the vertical axis denotes their percentage of variance explained. We notice that the percentage of variance explained does not increase significantly after 100 principle components.

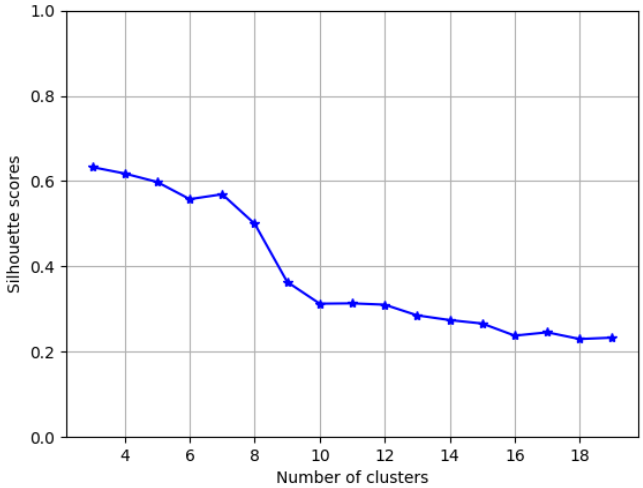

Figure 5: Silhouette scores for User Clustering.
The horizontal axis denotes the number of clusters chosen for K-means clustering, while the vertical axis denotes their the silhouette scores. We notice that the silhouette scores drop sharply after 7 clusters.

Tables

Table 1: Side effects of anti-depressants.

Drugs	Side effects
Lexapro	chills, constipation, cough, decreased appetite, **decreased sexual desire**, **diarrhea**, **dry mouth**, joint pain, muscle ache, tingling feeling, **sleepiness or unusual drowsiness**, unusual dream, **sweating**, ...
Xanax	abdominal or stomach pain, muscle weakness , **changed behavior**, chills, cough, decreased appetite, decreased urine, **diarrhea**, difficult bowel movement, cough, **dry mouth**, tingling feeling, **sleepiness or unusual drowsiness**, slurred speech, sweating, yellow eye...
Zoloft	**changed behavior**, decreased sexual desire, **diarrhea**, **dry mouth**, heartburn, **sleepiness or unusual drowsiness**, **sweating**...

The "Drugs" and "Side effects" columns respectively list the anti-depressants and their side effects extracted from a drug – side effect database. Side effects in common among those listed are bolded.

Table 2: A sample discussion thread from an online health community.

User IDs	Posts	Mentioned drugs	Aggregated side effects
3690	While my experience of 10 years is with Paxil, I expect that Zoloft will be the same. You should definitely feel better within 2 weeks. One way I found to make it easier to sleep was to get lots of exercize. Walk or run or whatever to burn off that anxiety.	Zoloft, Paxil	changed behavior, decreased sexual desire, diarrhea, dry mouth, heartburn, sleepiness or unusual drowsiness,...
26521	I've heard of people going "cold turkey" and having withdrawal at 6 months! Please, get in contact with a doctor ASAP! "common symptoms include dizziness, electric shock-like sensations, sweating, nausea, insomnia, tremor, confusion, nightmares and vertigo"		

The "User IDs" and "Posts" columns respectively list the IDs of users involved in the discussions and their messages. The "Mentioned drugs" and "Aggregated side effects" columns respectively list the explicitly discussed drugs and their combined side effects.

Table 3: Most common experienced side effects for each user cluster c_i ($i = 1$ to 7).

Cluster	Most common experienced side effects
c_1	vision blurred, yellow skin, vision double, yellow eye, nose stuffy
c_2	headache, itch, stomach pain, weak, nausea
c_3	itch, irritate, headache, pain abdominal, "stomach cramp"
c_4	bad taste, nausea, tiredness, irritate, mouth ulcer
c_5	skin red, itch, rash skin, skin peeling, "burning skin"
c_6	sneezing, nose runny, nose stuffy, decrease sexual desire, pain breast
c_7	nausea, stomach pain, vomit, diarrhea, pain abdominal

Table 4: Some statistics on our dataset.

# Users	14,966
# Threads	78,213
Avg. words per post	67.45
Avg. posts per thread	3.97
Avg. participated threads per user	54.7
# Side effects (SE)	315
Avg. SEs per thread	74.25
# Drugs	1869
Avg. experienced side effects per user	128.12

Table 5: Performance of CNN-based models and LSTM-based models in Ablation Study.

Systems	Components			Evaluation Metrics		
	CW	UE	CA	Pre.	Rec.	F_1
LSTM-Vanilla				0.6173	0.407	0.4335
LSTM-WPE	✓			**0.6376**	0.4344	0.4503
LSTM-WPEU	✓	✓		0.6064	0.5001	0.4896
LSTM-NEAT	✓	✓	✓	0.6197	**0.5134**	**0.5064**
CNN-Vanilla				0.7214	0.5503	0.5637
CNN-WPE	✓			**0.7423**	0.5799	0.5804
CNN-WPEU	✓	✓		0.6923	0.6350	0.5910
CNN-NEAT	✓	✓	✓	0.7066	**0.6431**	**0.6139**

In the "Components" column, "CW", "UE", "CA" denote "Credibility Weights", "User Expertise" and "Cluster Attention" module components, respectively. In the "Evaluation Metrics" column, "Pre.", "Rec." and "F_1" denote "Precision", "Recall", and "F_1 score".

Table 6: Analysis of NEAT's Credibility versus baselines in approximating credibility proxy.

Methods	Thread nDCG@2	Thread Spearman	Forum Spearman
Random	0.7968	-0.0271	0.0
Post frequency	0.8812	0.4223	0.1924
Question frequency	0.8341	0.1773	0.0279
NEAT's Credibility	**0.8856**	**0.4403**	**0.3055**

The "Thread nDCG@2", "Thread Spearman", and "Forum Spearman" columns respectively denote the values of Normalized Discounted Cumulative Gain at 2 at thread level, Spearman's rank correlation coefficient at thread level and forum level of each method when using rankings by number of "thanks" as ground truths.

Table 7: Performance of NEAT versus baselines in Side Effect Discovery of Ibuprofen, Levothyroxine, and Metoformin.

Methods	Ibuprofen			Levothyroxine			Metoformin		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1	Pre.	Rec.	F_1
RF	0.583	0.414	0.474	0.319	0.401	0.347	0.48	0.647	0.491
uNEAT	0.859	0.371	0.487	0.505	0.349	0.404	0.798	0.361	0.497
NEAT	0.845	0.427	**0.536**	0.549	0.385	**0.443**	0.814	0.365	**0.504**

In the "Methods" column, "RF" denotes Random Forest baseline from Bag-of-word, and "uNEAT" denotes User permutation baseline from NEAT. "Pre.", "Rec." and "F_1" denote "Precision", "Recall", and "F_1 score", respectively

Table 8: Performance of NEAT versus baselines in Side Effect Discovery of Omeprazole and Alprazolam.

Methods	Omeprazole			Alprazolam		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
RF	0.229	0.458	0.271	0.639	0.432	0.511
uNEAT	0.534	0.393	0.394	0.981	0.551	0.663
NEAT	0.522	0.421	**0.41**	0.977	0.596	**0.704**

In the "Methods" column, "RF" denotes Random Forest baseline from Bag-of-word, and "uNEAT" denotes User permutation baseline from NEAT. "Pre.", "Rec." and "F_1" denote "Precision", "Recall", and "F_1 score", respectively.

Table 9: Performance of CNN-NEAT's Attention versus baselines in Side Effect Extraction of Ibuprofen, Levothyroxine, Metoformin, Omeprazole and Alprazolam in term of Precision.

Methods	Ibuprofen	Levothyroxine	Metoformin	Omeprazole	Alprazolam
UMLS Tagging	0.6801	0.6145	0.8378	**0.5218**	0.614
Neural Extractor [19]	0.6741	0.6259	0.8092	0.4665	0.6161
CNN-NEAT's Attention	**0.7073**	**0.7119**	**0.8557**	0.504	**0.688**

Table 10: A test example highlighting the extracted side effects obtained by CNN-NEAT's Attention versus baselines.

User IDs	Cluster's side effects	Post content
8420	stomach pain, headache, itch, weak, nausea	I wont commit suicide but the discomfort <U> is enough to make me want to die right now [...] I feel like I have sever Akathisia (inner restlessness<U,X> that makes you feel like your body is electrified [...] I also have nausea<U,X,N>, but I can eat a little, sweating<U,X,N>/cold, and extreme fatigue<U,X>, although I already have chronic fatigue [...] I also get anxious<U,X,N> if I take more oxycodone for breakout pain<U> and then go right back<U> down

In the "Post content" column, the correct and wrong side effects are highlighted in blue and red. The extracted side effects of UMLS Tagging, Neural Extractor and CNN-NEAT's Attention are followed by $< U >$, $< X >$, and $< N >$, respectively. The "Cluster's side effects" column shows the list of common side effects in user 8420's cluster.

Table 11: Performance for individual integration of UE and CA in Ablation Study.

Systems	Components			Evaluation Metrics		
	CW	UE	CA	Pre.	Rec.	F_1
LSTM-UE		✓		0.6513	0.4204	0.4531
LSTM-CA			✓	0.6416	0.4293	0.4611
CNN-UE		✓		0.6738	0.6185	0.5743
CNN-CA			✓	0.7441	0.5616	0.5883

In the "Components" column, "CW", "UE", "CA" denote "Credibility Weights", "User Expertise" and "Cluster Attention" module components, respectively. In the "Evaluation Metrics" column, "Pre.", "Rec." and "F_1" denote "Precision", "Recall", and "F_1 score".