Predicting User Reported Drug Side Effects Using a Gated Neural Network

Lahari Poddar
School of Computing
National University of Singapore
lahari@comp.nus.edu.sg

Wynne Hsu
School of Computing
National University of Singapore
whsu@comp.nus.edu.sg

Mong Li Lee
School of Computing
National University of Singapore
leeml@comp.nus.edu.sg

ABSTRACT

The detection of Adverse Drug Events (ADE) or side effects of different drugs are necessary to minimize potential health risks of patients. Given the prevalence of user reported contents on Twitter and various health forums, recent research has focused on the automatic discovery of potential side effects of drugs from these online platforms. However, it is not clear if these reported side effects are solely due to the drug or there are other confounding factors that influence a patient’s experience of side effects. In this work, we seek to characterize the reported side effects along with their severity for different drugs, based on patients’ past drug evaluations and pre-existing medical conditions. We analyze a large dataset collected from an online health community patientslikeme.com, and observe that there exists a strong correlation between a patient’s existing health condition(s) and the side effects she reports across different drugs. We develop a multi-objective deep neural network architecture with gating mechanism, to predict the possible side effects and their overall severity level for a drug, for a given patient. Experimental results from a real world drug evaluations dataset demonstrate the effectiveness of our approach over state-of-the-art baselines. Furthermore, our adaptation of Mixture of Experts approach imbues the network with added explainability, allowing it to justify its predictions of side effects with respect to the evaluated drug, the patient and her conditions.

KEYWORDS
Computational Health, ADE Detection, Neural Network, Mixture of Experts, Side Effect Prediction, Personalization

ACM Reference Format:

1 INTRODUCTION

In the last decade, sharing of life events including medical and health information online has become widespread in both social networking sites (e.g., Twitter), and online health Forums (e.g., PatientsLikeMe, HealthBoards, WebMD). Online health communities constitute an important source of medical information, with 59% of the adult US population seeking health-related information from online resources [11], and nearly half of US physicians relying on them for professional use [10]. This has led to a surge of interest in the research community to automatically discover potential side effects of drugs from such online self-reported medical information [23, 29, 34, 41].

While these large amount of online health information can help us complement existing medical knowledge and speed up discoveries of potential drug reactions [12, 37, 38], there remains a widespread concern of whether the reported side effects are truly due to the drugs [6, 28]. Table 1 shows evaluations of two drugs by different patients from the patientslikeme forum. As we can see from the sample, different people experience different side effects with varying severity while using the same drug.

<table>
<thead>
<tr>
<th>Drug</th>
<th>User</th>
<th>Severity Rating</th>
<th>Reported Side Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clonazepam</td>
<td>u1</td>
<td>2</td>
<td>sleepiness</td>
</tr>
<tr>
<td></td>
<td>u2</td>
<td>1</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>u3</td>
<td>3</td>
<td>drowsiness, apathy, confusion, balance loss</td>
</tr>
<tr>
<td>Levothyroxine</td>
<td>u4</td>
<td>1</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>u5</td>
<td>3</td>
<td>weight gain</td>
</tr>
<tr>
<td></td>
<td>u6</td>
<td>3</td>
<td>weight gain, hair loss, quivering, insomnia</td>
</tr>
</tbody>
</table>

Table 1: Sample evaluations of two drugs by different patients from our dataset.

Our preliminary investigation shows that among the side effects reported by users, there exists a significant percentage of unsubstantiated\(^1\) side effects. Many of these side effects are, in fact, more correlated to the underlying medical condition(s) of the user than the drug for which they are reported. With more and more people joining the online health forums to seek information and support, it is important that these forums provide accurate information that is tailored according to individual user’ s condition, to prevent unnecessary anxiety [4, 35]. With this, we can reduce the number of users who might be reluctant to take a drug due to the long list of reported side effects, even though many of these reported side effects are not applicable to her. This motivates us to develop a framework that is able to better characterize the complex relationship between user—condition—drug and personalize the prediction of side effects for a specific user. Apart from the set of side effects, we also aim to predict a numeric score denoting the severity of the side effects for a user-drug interaction.

We formulate a multi-objective learner to predict both the set of side effects and the severity rating that a user reports for a drug. We design a novel deep neural network architecture called Multi

---

\(^1\)not associated with the drug as per expert medical knowledge base
objective Mixture of Experts (MoMEx) to encode the complex relationship between user–condition–drug combination and the target variable of side effects. MoMEx uses a gating network inspired from the mixture of experts model [16, 18]. It probabilistically combines the predictions from three local expert networks that are built to predict side effects based on user, condition, and drug separately. The gating network has an added advantage in that we are able to use the probabilities assigned to each of the local experts to explain why the model predicts a certain side effect that the user is likely to experience. This transparency is important, as it will enable a user to make a better decision concerning whether to take a prescribed drug or not.

To summarize, the main contributions of this work are:

- Systematically investigating the nature of user reported side effects in online health forums and discovering their correlation with the user and her pre-existing medical condition(s) apart from the drug;
- Proposing a multi-objective neural network architecture for predicting the side effects and their severity score based on the interaction between user, drug, and conditions;
- Demonstrating the effectiveness of MoMEx on a real world drug evaluation dataset, compared to state-of-the-art baselines and discuss the explainability of our predictions. This is followed by a review of relevant research works in Section 5. We conclude the paper and outline possible future directions in Section 6.

2 PRELIMINARIES

In this section, we first describe the dataset collected from an online health forum, highlighting the different signals and elements available. Thereafter we present an initial analysis of the dataset, that illustrates the challenges and motivates our approach.

2.1 Dataset

We constructed a dataset from a large online health community. We crawled PatientsLikeMe (PLM) in May 2017 for drug evaluations done by its users along with the associated users’ profile pages. Users can evaluate a drug based on four criteria, namely, Effectiveness, Side Effects, Adherence and Burden on a 5 point scale (ranging from 0 to 4), as shown in Figure 1. The rating for Side Effects contains a numeric score representing the severity and a list of side effects experienced after taking the drug. A snapshot of sample drug evaluations is shown in Figure 1.

A user can evaluate a drug multiple times for either different purposes or for having taken them at different times. Here, we only consider her most recent evaluation of a drug. We crawl all such evaluations for 1274 drugs as well as the profiles of the associated users. We filter out those side effects which have been mentioned less than 3 times in the whole dataset. In the user-profile, apart from gender, age and location, a user also mentions the pre-existing conditions she has been suffering from. We collect the set of medical conditions mentioned by a user in the profile page. Statistics of the PLM dataset are shown in Table 2.

![Figure 1: Sample drug evaluations by users. We focus on modeling only the reported side effects and the corresponding severity rating (highlighted in a blue rectangle).](image)

<table>
<thead>
<tr>
<th>Number of drugs</th>
<th>1274</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>8008</td>
</tr>
<tr>
<td>Number of unique conditions</td>
<td>1607</td>
</tr>
<tr>
<td>Number of unique side effects</td>
<td>1417</td>
</tr>
<tr>
<td>Number of evaluations</td>
<td>41,050</td>
</tr>
<tr>
<td>Average number of side effects in an evaluation</td>
<td>1.82</td>
</tr>
</tbody>
</table>

Table 2: Statistics of the PLM drug evaluation dataset.

2.2 Preliminary Study

To better understand the nature of user reported side effects in an open online setting, we carry out an initial study on our dataset. We aim to answer a few questions from this study.

Q1. Can all the user reported side effects be substantiated by an authoritative medical source?

We first compare the user reported side effects in our dataset with those published on the Mayo Clinic portal\(^2\) which contains curated expert information about drugs and their side effects. The side-effects for each drug are categorized into common, less common and rare.

For each drug mentioned in the PLM dataset, we obtain the set of all its reported side-effects across users. Then we match the drug to a drug-family in the Mayo Clinic portal and consider the corresponding listed side effects as the ground truth. We successfully match 679 out of 1274 drugs in the PLM dataset. Table 3 shows that only 61.53% of reported side effects are known common side effects of a drug, while 8.74% and 3.08% of the reported ones are less common and rare side effects respectively. This indicates that comparatively lesser known side effects of a drug are indeed reported.

\(^2\)mayoclinical.org/drugs-supplements/
in health forums and can help augment the existing medical knowledge. However, we also note that an alarming 26.65% of side effects reported for drugs in health forums do not match with any known side effects of the drug.

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common</td>
<td>61.53%</td>
</tr>
<tr>
<td>Less Common</td>
<td>8.74%</td>
</tr>
<tr>
<td>Rare</td>
<td>3.08%</td>
</tr>
<tr>
<td>Unsubstantiated</td>
<td>26.65%</td>
</tr>
</tbody>
</table>

Table 3: Percentage breakdown of reported side effects in the different categories.

Q2. Do patients report similar side effects across drugs?

We examine whether there are some side effects that users tend to experience and report regardless of the drug they take. We consider the set of all users who have evaluated at least 4 drugs. For each user, we compute a metric called recurring side effect ratio, which is defined as the fraction of side effects that has been reported in at least k drug evaluations by her. Assume a user has evaluated 4 drugs \( d_1, d_2, d_3, d_4 \) and reported the following side effects:

\[
\begin{array}{c|c}
\text{Drug} & \text{Side Effects} \\
\hline
\text{d}_1 & s_1, s_2, s_3, s_4 \\
\text{d}_2 & s_2, s_4 \\
\text{d}_3 & s_2, s_3, s_4 \\
\text{d}_4 & s_2, s_3 \\
\end{array}
\]

For \( k = 3 \), the side effects \( s_2, s_3 \) and \( s_4 \) are reported in at least 3 drug evaluations, hence the recurring side effect ratio is \( \frac{3}{1} \). For \( k = 4 \), only \( s_2 \) is reported in 4 drug evaluations, hence the recurring side effect ratio is \( \frac{1}{4} \). A high value for the ratio indicates that the user tends to report the same set of side effects regardless of the drug she takes, while a low ratio would indicate that she experiences different side effects for different drugs.

Q3. Do the pre-existing conditions of a patient have any correlation to the side-effects she reports across drugs?

We analyze whether the pre-existing conditions of a user influence the side effects she experiences and mistakenly reports as side effects of drugs. For example, a patient suffering from insomnia may experience fatigue or drowsiness, and report them as side effects of the drugs she is currently taking.

For each side effect in the PLM dataset, we consider its association with condition \( c \) and drug \( d \) respectively, using Jaccard similarity coefficient as,

\[
J(s, c) = \frac{\text{intersection}(U_s, U_c)}{\text{union}(U_s, U_c)} \\
J(s, d) = \frac{\text{intersection}(U_s, U_d)}{\text{union}(U_s, U_d)}
\]

We consider a side effect \( s \) is more correlated with a condition than a drug, if there exists a condition \( c \), for which \( J(s, c) > J(s, d) \) for all \( d \in D \), where \( D \) is the total number of drugs in the dataset. We find that around 74.02% of side-effects are more correlated with a condition than with a drug, indicating that the pre-existing conditions of a user are linked to the side effects reported.
3 PROPOSED MOMEX FRAMEWORK

Our preliminary study shows that the reported side effects are not solely due to the drugs, but might be correlated with some underlying medical conditions. This motivates us to propose an automatic approach towards predicting the side effects that a patient might report while evaluating a drug.

![Graphical representation of the interaction structure between user, her conditions, drugs, reported severity rating and side effects](image)

**Figure 4: A graphical representation of the interaction structure between user, her conditions, drugs, reported severity rating and side effects**

We present a graphical representation of the interactions in Figure 4. Users $u_1, u_2$ have pre-existing conditions $\{c_1, c_2, c_3\}$ and $\{c_2\}$ respectively. The most recent evaluation for drugs $d_1$ and $d_2$ by each user has a severity score as labeled on the edge between the user and drug. Associated with each edge is also the list of reported side effects among $s_1, s_2, s_3$.

We formulate the problem as a multi-objective prediction task. For a user $u$ and drug $d$, we predict the following:

- **Severity of Side effects**: a numerical rating $r_{ud}$, real-valued number in the range $[0, 4]$.
- **List of Side effects**: a sparse $S$ dimensional binary vector $s_{ud}$, indicating the side effects experienced by user $u$, for drug $d$, where $S$ is the total number of side effects.

We propose a neural network architecture, called MoMEx (Multi-objective Mixture of Experts), for predicting user-drug side effects along with their severity rating. The input signals to MoMEx are user, drug and condition, as depicted in Figure 5.

We use three separate embeddings to map these categorical inputs to a lower dimensional latent feature space of dimension $k$. Let $x_u$, $y_d$, $z_c$ denote the latent feature vectors of user $u$, drug $d$, and condition $c$ respectively.

Each user $u$ is associated with some conditions, denoted as $c_u$, a sparse binary vector of dimension $C$, where $C$ is the total number of distinct conditions. The $m^{th}$ entry in $c_u$ indicates whether the $m^{th}$ condition has been reported by user $u$. We embed each entry of $c_u$ using the condition embedding $z$ to get feature representation of individual conditions. We concatenate all the condition embeddings of a user and feed through a fully connected layer, to capture the dependencies between co-existing conditions. The output of this layer gives us a $k$ dimensional vector, $v_u$ representing the conditions of user $u$.

Given the representations of user, drug and conditions in terms of latent feature vectors, the following subsections describe how the two prediction tasks are performed.

3.1 Predicting Severity of Side Effects

Given the user and drug embeddings, we learn a model to predict the severity rating $r_{ud}$ by a user $u$ for a drug $d$. The severity rating contains characteristics of both the user and drug. To incorporate that, we first combine the latent features of the user and drug, by concatenating their embedding vectors $x_u$ and $y_d$. However, a straightforward concatenation is unable to capture the complex structure implied in the users’ historical interactions.

In order to learn the interactions between user and drug features, we add multiple fully connected layers on the concatenated vector. Through these hidden layers, we introduce flexibility and non-linearity in the model, to capture complex relationship between the user and drug interactions. The output of the $l^{th}$ hidden layer is given by:

$$h_l = \phi(W_l h_{l-1} + b_l)$$

where $W_l$ and $b_l$ are the weight matrix and bias vector for the $l^{th}$ layer respectively, $\phi$ is an activation function used to introduce non-linearity. We use $tanh$ activation function for our network to squash the input within the range $(-1, 1)$. In our experiments, we obtained comparable results for using $tanh$ and $Relu$ and slightly worse results for $sigmoid$, as activation functions.

The output to the first hidden layer is the concatenated vector of user and drug embeddings. The output of the last hidden layer is fed into an output layer, which transforms it to a real valued rating $r^*_{ud}$.

$$r^*_{ud} = \phi(W_l h_L + b_L)$$

We formulate this as a regression problem and the loss function is constructed as:

$$L^r = \sum_{(u,d) \in X} (r_{ud} - r^*_{ud})^2$$

where $X$ represents the training set, $r_{ud}$ represents the ground truth rating and $r^*_{ud}$ represents the predicted rating for drug $d$ by user $u$.

3.2 Predicting List of Side Effects

Now we proceed to describe our approach for predicting the list of side effects $s_{ud}$, the user $u$ reported for drug $d$. This list is a sparse binary vector where the $m^{th}$ entry indicates whether the $m^{th}$ side effect has been reported by the user for a drug. We consider this prediction task as multiple individual binary classifications, where the correlation among the labels is exploited by the latent space in our proposed neural network architecture.

Motivated by our initial study, we realize that the reported side effects of a drug $s_{ud}$, could be because of the drug $d$, or caused by the pre-existing conditions $c_u$ of the user $u$. Hence, we aim to learn a model that predicts $s_{ud}$ given the embeddings of user, drug and a user’s conditions i.e. $x_u$, $y_d$ and $v_u$ respectively.

A plausible approach could be concatenating all the three vectors and pass them through a multi-layer perceptron to get a binary prediction for each side effect. It would use non-linear transformations and combinations of user-drug-condition features to predict whether a side effect will occur. However, such a network will be unexplainable, and it will be difficult to rationalize why a certain side effect was predicted - whether it was because of the drug or...
condition of the user or a complicated non-linear combination of them.

Instead of building a joint model which takes the combination of user-drug-condition features as a single input, inspired by Mixture of Experts models, we develop three simpler local expert models namely, $E_{\text{drug}}$, $E_{\text{user}}$, and $E_{\text{cond}}$, taking as input the drug feature ($y_d$), user feature ($x_u$), and condition feature ($v_u$) respectively. The predictions from the local experts $E_{\text{drug}}$, $E_{\text{user}}$ and $E_{\text{cond}}$, are denoted as $\hat{s}_{\text{drug}}$, $\hat{s}_{\text{user}}$, and $\hat{s}_{\text{cond}}$ respectively. The $m^{th}$ entry of the vector $\hat{s}_{\text{drug}}$ denotes the probability of occurrence of the $m^{th}$ side effect according to the drug expert classifier.

Finally, we need to combine the predictions from these individual experts to output a single prediction $\hat{s}_{ud}$. One way of doing that could be just averaging their predictions, but that does not make sense in our case. When we average the output of multiple classifiers and try to match it to a target value, we force each of the classifier to compensate for the combined error made by the other classifiers. However, in our scenario, there are certain side effects that can be explained by only a single expert classifier (say, drug) and we can ignore the results of the other classifiers for that case. This motivates us to develop a gating function similar to Mixture of Experts model, where for each input we select an expert with some probability. The final prediction is a weighted average of the local predictions of the three classifiers and probabilities with which their predictions were combined. Therefore, if our network predicts that a certain side effect will be reported by a user for a drug, we can provide an explanation for the cause of the side effect with the probability distribution.

Now we need to define the structures of our expert networks and the gating network. We choose simple and similar structures for all three of our expert classifiers but with different parameters. For a user $u$ and drug $d$, the three classifiers $E_{\text{drug}}$, $E_{\text{user}}$ and $E_{\text{cond}}$, take the corresponding latent features $y_d$, $x_u$ and $v_u$, respectively, as input. The input is passed through one hidden layer before outputting the predictions $\hat{s}_{\text{drug}}$, $\hat{s}_{\text{user}}$, and $\hat{s}_{\text{cond}}$.

The gating network multiplies its input with a trainable weight matrix and applies a $\text{sigmoid}$ non-linearity to convert it to a vector of dimension $S$. This transforms the input from latent feature space to the side effect dimension. By multiplying the output of this layer, with a second trainable weight matrix, we transform the value in each side effect dimension, to a 3-dimensional vector representing the weights for each of the three experts. A $\text{softmax}$ activation function is then applied on each of these vectors, to convert its elements to real values in the range $[0, 1]$ that add up to 1.

We train the gating network by back-propagation, along with the rest of the model. Gradients are also back-propagated through the gating network to its inputs. Following the Mixture of Experts model, we develop three simpler local expert models with drug, user, and condition experts respectively. The final prediction is a weighted average of the local predictions of the three classifiers and probabilities with which their predictions were combined. Therefore, if our network predicts that a certain side effect will be reported by a user for a drug, we can provide an explanation for the cause of the side effect with the probability distribution.

Now we need to define the structures of our expert networks and the gating network. We choose simple and similar structures for all three of our expert classifiers but with different parameters. For a user $u$ and drug $d$, the three classifiers $E_{\text{drug}}$, $E_{\text{user}}$ and $E_{\text{cond}}$, take the corresponding latent features $y_d$, $x_u$ and $v_u$, respectively, as input. The input is passed through one hidden layer before outputting the predictions $\hat{s}_{\text{drug}}$, $\hat{s}_{\text{user}}$, and $\hat{s}_{\text{cond}}$.

The gating network multiplies its input with a trainable weight matrix and applies a $\text{sigmoid}$ non-linearity to convert it to a vector of dimension $S$. This transforms the input from latent feature space to the side effect dimension. By multiplying the output of this layer, with a second trainable weight matrix, we transform the value in each side effect dimension, to a 3-dimensional vector representing the weights for each of the three experts. A $\text{softmax}$ activation function is then applied on each of these vectors, to convert its elements to real values in the range $[0, 1]$ that add up to 1.

We train the gating network by back-propagation, along with the rest of the model. Gradients are also back-propagated through the gating network to its inputs. Following the Mixture of Experts model, we develop three simpler local expert models with drug, user, and condition experts respectively. The final prediction is a weighted average of the local predictions of the three classifiers and probabilities with which their predictions were combined. Therefore, if our network predicts that a certain side effect will be reported by a user for a drug, we can provide an explanation for the cause of the side effect with the probability distribution.

Now we need to define the structures of our expert networks and the gating network. We choose simple and similar structures for all three of our expert classifiers but with different parameters. For a user $u$ and drug $d$, the three classifiers $E_{\text{drug}}$, $E_{\text{user}}$ and $E_{\text{cond}}$, take the corresponding latent features $y_d$, $x_u$ and $v_u$, respectively, as input. The input is passed through one hidden layer before outputting the predictions $\hat{s}_{\text{drug}}$, $\hat{s}_{\text{user}}$, and $\hat{s}_{\text{cond}}$.

The gating network multiplies its input with a trainable weight matrix and applies a $\text{sigmoid}$ non-linearity to convert it to a vector of dimension $S$. This transforms the input from latent feature space to the side effect dimension. By multiplying the output of this layer, with a second trainable weight matrix, we transform the value in each side effect dimension, to a 3-dimensional vector representing the weights for each of the three experts. A $\text{softmax}$ activation function is then applied on each of these vectors, to convert its elements to real values in the range $[0, 1]$ that add up to 1.

We train the gating network by back-propagation, along with the rest of the model. Gradients are also back-propagated through the gating network to its inputs. Following the Mixture of Experts model, we develop three simpler local expert models with drug, user, and condition experts respectively. The final prediction is a weighted average of the local predictions of the three classifiers and probabilities with which their predictions were combined. Therefore, if our network predicts that a certain side effect will be reported by a user for a drug, we can provide an explanation for the cause of the side effect with the probability distribution.

Now we need to define the structures of our expert networks and the gating network. We choose simple and similar structures for all three of our expert classifiers but with different parameters. For a user $u$ and drug $d$, the three classifiers $E_{\text{drug}}$, $E_{\text{user}}$ and $E_{\text{cond}}$, take the corresponding latent features $y_d$, $x_u$ and $v_u$, respectively, as input. The input is passed through one hidden layer before outputting the predictions $\hat{s}_{\text{drug}}$, $\hat{s}_{\text{user}}$, and $\hat{s}_{\text{cond}}$.

The gating network multiplies its input with a trainable weight matrix and applies a $\text{sigmoid}$ non-linearity to convert it to a vector of dimension $S$. This transforms the input from latent feature space to the side effect dimension. By multiplying the output of this layer, with a second trainable weight matrix, we transform the value in each side effect dimension, to a 3-dimensional vector representing the weights for each of the three experts. A $\text{softmax}$ activation function is then applied on each of these vectors, to convert its elements to real values in the range $[0, 1]$ that add up to 1.

We train the gating network by back-propagation, along with the rest of the model. Gradients are also back-propagated through the gating network to its inputs. Following the Mixture of Experts model, we develop three simpler local expert models with drug, user, and condition experts respectively. The final prediction is a weighted average of the local predictions of the three classifiers and probabilities with which their predictions were combined. Therefore, if our network predicts that a certain side effect will be reported by a user for a drug, we can provide an explanation for the cause of the side effect with the probability distribution.

Now we need to define the structures of our expert networks and the gating network. We choose simple and similar structures for all three of our expert classifiers but with different parameters. For a user $u$ and drug $d$, the three classifiers $E_{\text{drug}}$, $E_{\text{user}}$ and $E_{\text{cond}}$, take the corresponding latent features $y_d$, $x_u$ and $v_u$, respectively, as input. The input is passed through one hidden layer before outputting the predictions $\hat{s}_{\text{drug}}$, $\hat{s}_{\text{user}}$, and $\hat{s}_{\text{cond}}$.

The gating network multiplies its input with a trainable weight matrix and applies a $\text{sigmoid}$ non-linearity to convert it to a vector of dimension $S$. This transforms the input from latent feature space to the side effect dimension. By multiplying the output of this layer, with a second trainable weight matrix, we transform the value in each side effect dimension, to a 3-dimensional vector representing the weights for each of the three experts. A $\text{softmax}$ activation function is then applied on each of these vectors, to convert its elements to real values in the range $[0, 1]$ that add up to 1.

We train the gating network by back-propagation, along with the rest of the model. Gradients are also back-propagated through the gating network to its inputs. Following the Mixture of Experts model, we develop three simpler local expert models with drug, user, and condition experts respectively. The final prediction is a weighted average of the local predictions of the three classifiers and probabilities with which their predictions were combined. Therefore, if our network predicts that a certain side effect will be reported by a user for a drug, we can provide an explanation for the cause of the side effect with the probability distribution.

Now we need to define the structures of our expert networks and the gating network. We choose simple and similar structures for all three of our expert classifiers but with different parameters. For a user $u$ and drug $d$, the three classifiers $E_{\text{drug}}$, $E_{\text{user}}$ and $E_{\text{cond}}$, take the corresponding latent features $y_d$, $x_u$ and $v_u$, respectively, as input. The input is passed through one hidden layer before outputting the predictions $\hat{s}_{\text{drug}}$, $\hat{s}_{\text{user}}$, and $\hat{s}_{\text{cond}}$.

The gating network multiplies its input with a trainable weight matrix and applies a $\text{sigmoid}$ non-linearity to convert it to a vector of dimension $S$. This transforms the input from latent feature space to the side effect dimension. By multiplying the output of this layer, with a second trainable weight matrix, we transform the value in each side effect dimension, to a 3-dimensional vector representing the weights for each of the three experts. A $\text{softmax}$ activation function is then applied on each of these vectors, to convert its elements to real values in the range $[0, 1]$ that add up to 1.
paradigm, we define the loss function for the side effect prediction module as

\[ L^s = \sum_{(u,d) \in X} \left( w_{ud}^{\text{user}} \cdot \text{Binary\_Crossentropy}(s_{ud}, \hat{s}_{ud}^{\text{user}}) \\ + w_{ud}^{\text{drug}} \cdot \text{Binary\_Crossentropy}(s_{ud}, \hat{s}_{ud}^{\text{drug}}) \\ + w_{ud}^{\text{cond}} \cdot \text{Binary\_Crossentropy}(s_{ud}, \hat{s}_{ud}^{\text{cond}}) \right) \]  

Binary\_Crossentropy(t, o) = − \{t \cdot \log(o) + (1 - t) \cdot \log(1 - o)\}  

where X represents the training set, s_{ud} represents the ground truth side effect vector of drug d by user u. In the Binary\_Crossentropy equation, t is the target value and o is the output value.

A loss function like this will encourage specialization, since we are comparing the prediction of each expert separately with the target and then training to reduce the weighted average of all these discrepancies, where the weights are the probabilities of selecting the experts through the gating network.

3.3 Multi-Objective Learning

We integrate both the prediction tasks into a unified multi-objective framework. The loss function for the entire framework is given by a weighted summation of the losses of its components

\[ L = \sum_{(u,d) \in X} \lambda_r L^r + \lambda_s L^s \]  

where \( L^r \) and \( L^s \) are the losses for severity rating prediction and side effects prediction respectively and \( \lambda_r \) and \( \lambda_s \) are the weights for the corresponding losses. In our experiments, we set them to be equal. One could vary them depending on which prediction task is more important. The whole network is trained using back-propagation in an end-to-end paradigm.

4 EVALUATION

We design our experiments to evaluate the effectiveness of the proposed framework in the following aspects:

- **Prediction of Severity Rating** (Section 4.2): Evaluating the performance of the proposed model for severity rating prediction compared to state-of-the-art methods.
- **Prediction of Side Effects** (Section 4.3): Evaluating the accuracy with which our model can predict the user reported side effects of a drug.
- **Explainability** (Section 4.4): Analyzing the helpfulness of the local expert models and gating layer for explaining the possible causes of the side effects.

4.1 Experimental Settings

We use the PLM dataset (described in Section 2.1) for our experiments. We divide it into training (80%), validation (10%) and test (10%) sets using five fold cross validation. All the reported results are obtained by averaging the results over five folds. All the parameters of our model as well as the competitive models are tuned via grid search on the validation set. The dimension of the embedding vectors (for user, drug and condition) are set to 64 unless otherwise mentioned. The number of fully connected layers for user-condition vector is 1 and is 3 for the rating predictor component. The number of fully connected layers in the local expert models and gating network are 1 and 2 respectively (increasing the number of layers led to overfitting in our dataset). Number of neurons in the fully connected layers of the rating predictor component and gating network are 128 and 256 respectively. We randomly initialized all model parameters with a Gaussian distribution (with mean 0 and standard deviation 0.01). The batch size for mini-batch training is 512 and the network is optimized using Adam[20] optimizer. The learning rate is set to 0.001. The proposed model is implemented using Keras[5] with Theano [2] as the backend.

4.2 Severity Rating Prediction

We start with evaluating our model on the severity rating prediction task using five-fold cross validation method. For each user-drug pair in the test set, we predict the severity rating and measure the prediction error using the most commonly used metrics, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

We compare our model, MoMEx, with the following state-of-the-art rating prediction models using the librec\(^1\) package.

- **NMF**: [22] Non-negative matrix factorization is a classic and one of the most widely used collaborative filtering approaches to the rating prediction task.
- **PMF**: [25] Probabilistic Matrix Factorization (PMF) can scale linearly with the number of observations and performs well for sparse and large datasets.
- **BPMF**: [33]: Bayesian Probabilistic Matrix Factorization (BPMF) is a fully Bayesian approach of solving Probabilistic Matrix Factorization (PMF) [25]. BPMF has been shown to achieve higher accuracy than PMF on some datasets.
- **SVD++**: [21]: A model merging the two approaches of collaborative filtering - latent factor models and neighbourhood models. It is a highly competitive baseline for rating prediction.

The number of latent factor is an important parameter in determining the capability of a factorization model. We vary the number of latent factors in the range \{4,8,16,32,64\} and compute accuracy for each competing model. For our neural network model, the dimensions of latent user, drug vectors are similar in spirit with the latent factors of a CF model and determine the predictive capability of our model [15]. Therefore we vary this dimension within the specified range and compare with other methods.

Figure 6 shows the performance of different models measured using MAE and RMSE with varying number of latent factors. We see that our method consistently achieves the best performance, across the varying factors, outperforming state-of-the-art rating prediction models. It outperforms the second best method SVD++ with a 4.11\% and 3.42\% improvement on an average, in terms of MAE and RMSE respectively. Furthermore, our method is more robust to variations in number of latent factors and the improvement over baselines is significantly higher, 4.74\% and 6.14\% for MAE and RMSE respectively, when number of latent factors is 4.

---

1[https://www.librec.net/](https://www.librec.net/)
4.3 Side Effect Prediction

Next, we evaluate the prediction of the list of side effects for a user given a drug. This is a challenging task as the class distribution is highly skewed. For a given user-drug interaction, only a few side effects are reported among a huge list of side effects. We use the standard precision, recall, and F1-score of the positive class (i.e. of the reported side effects) as evaluation metrics. To understand the contribution of each of these signals, we perform an ablation study with our MoMEx model. Table 4 shows the results. Unsurprisingly, MoMEx achieves the best F1-score when it takes into consideration all the three input signals, instead of taking a subset of them. This proves the necessity of modeling all the three contributing factors in side effect reporting.

We also compare MoMEx with the following variants:

- **Multi-Objective Multi Layer Perceptron (MoMLP):** We replace the mixture of experts network from our model with the standard Multi Layer Perceptron in this variation. We concatenate the user-, drug-, condition-latent vectors and feed them to the MLP layer and predict the list of side effects. We experimented with 1 – 3 number of fully connected layers for the MLP, and reported the best results.

- **Single Objective Mixture of Experts (SoMEx):** Instead of predicting both the severity rating and the side effects in the same model, we predict only the side effects using a single objective loss function for this baseline.

Table 5 shows that using the mixture of experts gives superior performance compared to Multi-Layer Perceptron. Furthermore, using a single objective loss function results in a slightly worse performance compared to MoMEx. This indicates that the joint modeling of both the severity rating and side effects using multi-objective learning benefits the side effect prediction task, as both of them essentially constitute a single evaluation of a drug by a user. When a user gives a severity rating of 0, we learn that the side effect experienced by this user is likely to be nil. On the other hand, when a user gives a high severity rating, the list of side effects to be predicted is likely to be long.

```
<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoMLEx (user + drug + condition)</td>
<td>0.612</td>
<td>0.457</td>
<td>0.525</td>
</tr>
<tr>
<td>SoMEx (user + drug + condition)</td>
<td>0.586</td>
<td>0.479</td>
<td>0.527</td>
</tr>
<tr>
<td>MoMEx (user + drug + condition)</td>
<td>0.615</td>
<td>0.519</td>
<td>0.563</td>
</tr>
</tbody>
</table>
```

Table 5: Performance comparison among MoMEx Variants

Finally, we compare MoMEx with AdaBoost, K Nearest Neighbour, and Random Forest classifiers using the implementations in scikit-learn python package [27]. Table 6 shows the performance comparison results. MoMEx clearly outperforms all the other methods across all the metrics. K Nearest Neighbour classifier suffers the most due to the highly skewed distribution. In general, these methods achieve a comparable precision but the recall is lower as they tend to miss many of the actual side effects in their prediction. MoMEx is able to exploit the correlation between side effects using the weights of the shared hidden layers and hence can achieve the best scores.

```
<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ada Boost</td>
<td>0.589</td>
<td>0.460</td>
<td>0.517</td>
</tr>
<tr>
<td>K Nearest Neighbor</td>
<td>0.151</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.611</td>
<td>0.439</td>
<td>0.511</td>
</tr>
<tr>
<td>MoMEx (user + drug + condition)</td>
<td>0.615</td>
<td>0.519</td>
<td>0.563</td>
</tr>
</tbody>
</table>
```

Table 6: Performance comparison with state-of-the-art classifiers

4.4 Explainability of MoMEx

A major advantage of using a mixture of experts framework for side effect prediction is that the gating network outputs a probability distribution over the local experts, \(E_{\text{drug}}, E_{\text{cond}}\) and \(E_{\text{user}}\). Recall \(E_{\text{drug}}, E_{\text{cond}}\) and \(E_{\text{user}}\) are the local expert models built using the drug, condition and user signals respectively. The prediction of a side effect for a given user is based on the weighted probabilities of these local experts. Armed with this probability distribution, we attempt to provide an explanation for the predicted side effects.

First, we characterize the difference between the probability distributions of the set of substantiated versus unsubstantiated side effects. Figure 7 shows the average probability with which the predictions of the local expert models are weighted to generate the final prediction for the set of substantiated side effects versus the set of unsubstantiated side effects. We focus only on \(E_{\text{drug}}\) and \(E_{\text{cond}}\). This is because the number of data points available for each
individual user is quite sparse (the lowest evaluation count for a user is 1, and the largest being 61). We observe that the weights for $E_{\text{user}}$ vary greatly depending on the user’s evaluation count, therefore the weights of $E_{\text{user}}$ would be difficult to generalize across users.

![Figure 7](image)

**Figure 7:** Probabilities assigned by the gating network to $E_{\text{drug}}$ and $E_{\text{cond}}$ for substantiated vs. unsubstantiated side effects

From Figure 7, we first note that for both sets, the probabilities assigned to $E_{\text{cond}}$ are higher than $E_{\text{drug}}$. This is consistent with our observation in Section 2.2 that many of the side effects are correlated to users’ medical conditions rather than to the drug they are taking. Additionally, this correlation seems to be more pronounced in the set of unsubstantiated side effects compared to the set of substantiated side effects.

Looking at the weights assigned to $E_{\text{drug}}$ for the two sets of side effects (see Figure 7a), we observe that a higher probability is assigned to the set of substantiated side effects compared to the set of unsubstantiated ones. This is intuitive, since the side effects that are known to be associated with a drug, will be reported by many users of the drug and $E_{\text{drug}}$ will be able to predict them reliably. On the other hand, side effects that are not substantiated will rarely be reported by users, resulting in $E_{\text{drug}}$ being unable to model it, and it will be assigned a lower weight by the gating network. Interestingly, the opposite phenomenon is observed for $E_{\text{cond}}$ (see Figure 7b). The unsubstantiated side effects receive relatively higher weight by the gating network compared to the substantiated ones. This indicates that users report some side effects that are not associated with the evaluated drugs, but rather, with their medical conditions. Such side effects are more reliably predicted by $E_{\text{cond}}$.

Table 7 shows a sample of the side effects correctly predicted by MoMEx and the corresponding probabilities of the local experts. The local expert with the highest probability (highlighted in bold) is the likely reason for the user to experience the predicted side effect. We observe that most of substantiated side effects correspond to $E_{\text{drug}}$, indicating that the side effects are due to the drug. In contrast, the unsubstantiated side effects (in italics) correspond to $E_{\text{cond}}$, suggesting that they are likely to be symptoms of users’ pre-existing conditions.

For example, user1 reports ‘loss of appetite’ and ‘sleeplessness’ after taking the drug Bupropion. MoMEx correctly predicts these side effects for user1 and further explains that the ‘loss of appetite’ is likely to be caused by Bupropion, while ‘sleeplessness’ is a result of her underlying conditions. We also note that user1 reports a substantiated side effect of ‘headache’ for the drug Fludrocortisone. MoMEx is able to make this prediction and attribute this to be possibly due to both the drug and one of her condition Migraine.

In another example, user2 suffers from insomnia and she has reported ‘sleeplessness’ across different drugs. MoMEx is able to explain that this side effect is not due to any of the drugs she takes but her pre-existing condition.

These cases demonstrate that analyzing the probability distributions of local experts generated from large scale user data is useful in interpreting the reported side effects, and could be of interest to both the web mining and medical communities.

5 RELATED WORK

**ADE detection:** Pharmaceutical companies carry out laboratory clinical trials and post-market surveillance, to discover side effects of drugs. However they are either limited in number or incur significant time delays to gather enough information [7, 31]. Given the large volumes of online content, recent research has focused on augmenting knowledge bases by discovering ADEs faster. Most of these research works address this as a supervised [8, 17, 26, 29, 34, 41] or semi-supervised [14, 23] binary classification where the objective is to detect mentions of ADE from post texts.

While detecting ADE mentions from online posts is an important task, the effects of other possible confounding factors are not captured there. Since different people may experience different side effects for the same drug, it is not sufficient to generate a single list of possible ADEs for a drug ignoring the user characteristics. Our work focuses on the personalized aspects of side effects reporting of different users and therefore is more closer to a recommendation system, where we try to predict the reaction (side effects and their severity) of a user (patient) to an item (drug).

**Recommendation Systems:** Collaborative Filtering based approaches have been widely used for recommendation systems in the past decade. Matrix Factorization based methods [21, 22, 25, 33] map users and items into a shared latent feature space and compute the inner product of their latent vectors to reflect the interactions between users and items i.e. the ratings. However they can only model a linear relationship using the dot product which limits their expressive power.

Recently, few neural network based architectures [15, 39] have been proposed to model the non-linear interaction between user-item features. They use multiple stacked hidden layers with non-linear activation function to capture more abstract features from user-item interactions. However these models focus on implicit user feedback for item recommendation instead of rating prediction, which is our objective in this work. We do not wish to recommend a drug to a user but aim to predict the possible side effects and their severity, if a user suffering from certain conditions, consumes a drug. For such explicit rating prediction, a recent work has proposed a neural network that uses not only the user-item information but also the review text accompanying the rating [24]. However we do not have such textual reviews available in our dataset and only focus on user-drug interactions to predict the rating.

There is another relevant class of recommendation algorithms which are context aware e.g., Tensor Factorization, Factorization Machines [19, 32]. Along with the usual user-item interactions, they
6 CONCLUSION

In this paper we have systematically investigated the characteristics of user reported side effects in online health forums. In our study conducted on a dataset collected from a large online health forum consisting of drug evaluations, we find that there is a significant percentage of unsubstantiated side effects reported for a drug. We also find that the reported side effects are more correlated with a user’s pre-existing medical conditions than with the drug. This motivated us to view the side effect detection problem as a personalized recommendation task instead of the heavily researched information extraction task.

We have proposed a novel neural network architecture to predict user responses to a drug in terms of side effects and severity rating while uncovering the possible factors behind the reported side effects. In order to predict both the severity rating and side effects together, we have designed a multi-objective learner with a combined loss function for the two prediction tasks. Our MoMEx framework is inspired from Mixture of Experts approach to train local experts based on user, condition and drug, and thereafter probabilistically combine their predictions using a gating layer.

Experimental evaluations done on a real world dataset shows that our approach is able to correctly model the factors influencing a side effect and outperforms state-of-the-art approaches on both prediction tasks. Additionally, our model learns a probability distribution over the local experts depending on their prediction accuracy for the side effect, which can be interpreted as the likeliness of the side effect being influenced by that factor.

Future work includes considering user demographics (e.g. age, gender etc) as well as concurrent drug uses which could also potentially influence how a user reacts to a drug, and multiple evaluations of the same drug by a user at different timestamps. Modeling a user’s temporal behavior could also reveal shifts in her drug reactions due to possible change of demography or conditions. More importantly, although we specialize our model for the use of side effect prediction in this paper, we believe our model is general in nature and could be applicable to other scenarios involving users, items and multiple interaction targets.

ACKNOWLEDGMENTS

We thank Divyat Mahajan from IIT Kanpur for helping us in collection and processing of the dataset during his internship in NUS. We would also like to thank the reviewers for their insightful comments.
REFERENCES


