Determining Associations with Word Embedding in Heterogeneous Network for Detecting Off-label Drug Uses

Christopher C. Yang
College of Computing and Informatics
Drexel University
Philadelphia, PA, US
chris.yang@drexel.edu

Mengnan Zhao
College of Computing and Informatics
Drexel University
Philadelphia, PA, US
mz438@drexel.edu

***Abstract*: Off-label drug use is quite common in clinical practice and inevitable to some extent. Such uses might deliver effective treatment and suggest clinical innovation sometimes, however, they have the unknown risk to cause serious outcomes due to lacking scientific support. As gaining information about off-label drug use could present a clue to the stakeholders such as healthcare professionals and medication manufacturers to further the investigation on drug efficacy and safety, it raises the need to develop a systematic way to detect off-label drug uses. Considering the increasing discussions in online health communities (OHCs) among the health consumers, we proposed to harness the large volume of timely information in OHCs to develop an automated method for detecting off-label drug uses from health consumer generated data. From the text corpus, we extracted medical entities (diseases, drugs, and adverse drug reactions) with lexicon-based approaches and measured their interactions with word embedding models, based on which, we constructed a heterogeneous healthcare network. We defined several meta-path-based indicators to describe the drug-disease associations in the heterogeneous network and used them as features to train a binary classifier built on Random Forest algorithm, to recognize the known drug-disease associations. The classification model obtained better results when incorporating word embedding features and achieved the best performance when using both association rule mining features and word embedding features, with F1-score reaching 0.939, based on which, we identified 2,125 possible off-label drug uses and checked their potential by searching evidence in PubMed and FAERS.**

Keywords: off-label drug use; online health community; word embedding; heterogeneous network

I. INTRODUCTION

Off-label drug use refers to prescribing marketed medications for indications that are not on their labeling information approved by Food and Drug Administration (FDA). Although FDA manages what medications can be brought to the market, they do not control or supervise whether the drugs are prescribed for their FDA-approved indications [1].

In clinical practice, off-label drug uses are very common and some off-label drugs have become a widely accepted practice for the disease. For instance, psychiatric drugs approved for one psychiatric disorder are often used for the other psychiatric conditions. In addition, off-label drug uses tend to occur more often for specific populations such as children, pregnant women, and the elderly, because they are often excluded for medication tests or clinical trials and there are less medications studied and approved for them specially [2]. Off-label drug uses occur highly in life-threatening and terminal conditions as well, when the physicians would like to give any treatment that might be useful no matter whether the medications are on-label or off-label.

Although physicians prescribe off-label drugs for the benefits of patients, and sometimes they suggest a possible clinical innovation, however, such uses still have a high risk to cause serious outcomes, adverse effects, or medication errors, due to the lack of scientific evidence. As off-label drug use seems to be inevitable to a great extent, it puts forward the need for a systematic way to identify off-label drug uses, which could present a clue to the stakeholders for further investigation on the medication efficacy and safety [2]. Thus, healthcare providers and patients can gain information about off-label uses in practice timely, and biomedical researchers can use the existing, especially those successful, off-label practices to assess the potential benefits and risks.

Survey is a popular approach to detect off-label uses currently, but limited by multiple conditions such as the number of respondents, the quality and truthfulness of answers, and the cost. Recently, the document from medication providers such as electronic health records (EHRs) and clinical notes provide a resource to detect off-label drug use in a scalable manner [3], meanwhile, large volumes of data generated by medication receivers, such as posts and comments on social media and online health communities (OHCs), also offer a great resource for detecting off-label uses automatically.

In this work, we developed an automated method to detect off-label drug uses from health consumer contributed data on an OHC website. With data coming from the OHC website, we first preprocessed the text and extracted the most common three medical entities mentioned in health consumer contributed contents: disease, drug, and adverse drug reaction (ADR). We developed two approaches to represent the interactions between the entities: (1) measuring their co-occurrence frequency from the population level and (2) using the state-of-the-art NLP algorithm-word embedding. We used the approach of co-occurrence frequency as a benchmark. Word embedding refers to the techniques of representing words with low-dimensional real-valued vectors by involving the contextual information of words during the computation. The generated vectors become meaningful representations and mirror the relationships between words. Word embedding has

978-1-5090-4881-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ICHI.2017.78

been successfully applied to many NLP tasks as well as in biomedical field (e.g. drug-drug interaction). We exploited this technique to analyze the interactions between medical entities. With the extracted medical entities and interactions, we constructed a heterogeneous healthcare network, on which, we determined the meta-paths between drugs and diseases and defined several meta-path-based indicators. We utilized those indicators as features to train a binary classifier to recognize the known drug-disease associations and to predict the possible off-label drug-disease relationships.

II. RELATED WORK

A. Off-label Drug Use Detection

Survey is the most popular research approach of exploring the issues associated with off-label drug use [4-6]. For instance, Conroy et al. [4] designed a prospective study to observe the off-label drug use for children in European countries and found that 39% of the prescriptions included off-label medications. Some studies investigate the associations between ADRs and off-label drug uses [7-10], and found that the percentage of ADRs associated with off-label prescriptions were distinctly higher than associated with licensed prescriptions, therefore, ADR problems should be closely supervised after prescribing off-label drugs.

With the digital availability of biomedical documents, some studies developed automated approaches to infer novel drug-disease associations to discover off-label uses. Jung et al. [11] focused on detecting off-label drug uses from the free-text clinical notes. Since a large number of off-label related articles are included in Excerpta Medica Database (EMBASE), but not labeled explicitly, Mesgarpour et al. [12] focused on developing highly sensitive search queries to retrieve off-label documents.

B. Prediction of Drug-Disease Relationship

Predicting the associations between drugs and diseases is the most critical step in detecting off-label drug-disease pairs and has been researched in many previous studies with methods derived from text mining and natural language processing (NLP). Xu & Wang [13] developed a lexicon-based approach to extract drug-disease treatment relationships from MEDLINE literature. Gottlieb et al. [14] calculated and ranked the similarity between drug-disease pairs based on the assumption that similar drugs are indicated from similar diseases. Besides NLP techniques, some studies unveil the drug-disease associations by using heterogeneous network-based methods such as iterative propagating algorithm [15], propagation flow algorithm [16-17], and Random Walk [18].

C. Word Embedding in Biomedical Informatics

Word embedding, referring to the techniques of representing words with vector space models (VSM), has been a major focus in NLP since its appearance. In the biomedical domain, word embeddings are mostly used for biomedical named entity recognition (BNER) and the evaluations are usually conducted on several popular BNER research tasks such as JNLPBA [19], BioCreAtIvE [20], and BioNLP Shared Tasks. Moreover, Wang et al. [21] used the feature vectors obtained by word embedding model for bio-event trigger detection and achieved a micro-averaging F1 score of 78.27%. Li et al. [22] incorporated word embedding features with bag-of-words (BOW) features for bio-event extraction and obtained the best performance using combined features.

III. OFF-LABEL DRUG USE DETECTION USING HETEROGENEOUS NETWORK MINING

The previous research presents a promising way to discover drug-disease associations by heterogeneous network-based methods, based on which, we developed an automated approach to detect off-label drug uses from user generated content in OHCs. Firstly, we collected health consumer contributed data from the OHC website and preprocessed the raw data to texts that only contain posts and comments. From the free-text corpus, we detected the medical entities with lexicon based approaches and extracted the interactions between entities by measuring their co-occurrence or cosine similarity in the vector space. Secondly, we constructed a heterogeneous network with the extracted medical entities and interactions, containing three types of nodes (diseases, drugs, and ADRs) and six types of edges (disease-disease, drug-drug, ADR-ADR, disease-drug, disease-ADR, and drug-ADR), and determined the meta-paths. Thirdly, we defined three meta-path-based indicators to describe the associations between drugs and diseases in the network and used them as input features to train a classifier that distinguishes known disease-drug pairs with those unknown. Lastly, we identified the off-label drug uses from the unknown pairs.

A. Dataset and Preprocessing

A variety of resources has been used to unveil drug-disease associations and detect off-label drug uses, mainly including pharmaceutical databases, biomedical literature, clinical text, and EHRs. Besides the resources provided by healthcare professionals, healthcare consumers generate large volumes of data by themselves as well, especially with the development of Web 2.0. In recent year, not only social media websites like Facebook and Twitter but also OHCs like MedHelp and PatientsLikeMe attract a large number of online users across the world.

1) Detection of medical entities

Unlike the biomedical databases that may include multiple medical entities such as gene, protein, and compound, OHC data are mostly contributed by consumers without professional background in medicine discussing their diseases, the drugs they take, and the side effects they have. Therefore, we only involved three medical entities in the work: disease, drug, and ADR. We applied lexicon-based approaches to detect diseases, drugs, and ADRs from the text corpus.

We resorted to UMLS, DrugBank, PharmGKB to build two lexicons of diseases and drugs, and tagged them with all the suggested names in the lexicon. In addition, considering that the word embedding model calculates vectors for each single word rather than the phrase and the common solution is to replace “x y” with “x_y”, we replaced the tagged entities with

their UMLS-id in the text corpus to guarantee they are represented by single words.

Compared with the detection of diseases and drugs, the detection of ADRs is more complicated, because consumers usually describe their adverse reactions or conditions with various and diverse expressions. Therefore, the standard medical databases are not appropriate for tagging ADRs from consumer contributed data. To deal with this problem, we employed Consumer Health Vocabulary (CHV) Wiki to build our ADR lexicon [23]. CHV Wiki integrated the everyday expressions of healthcare issues with the professional expressions [24]. Specifically, for each ADR, it provides its UMLS-id, its preferred name in UMLS, and common expressions by consumers. We used all the expressions in CHV Wiki to tagged ADRs in the corpus and then replaced them with their corresponding UMLS-id.

2) Detection of interactions between medical entities

There were three types of medical entities involved here: disease, drug, and ADR, thus there were six types of connections between them: disease-disease, drug-drug, ADR-ADR, disease-drug, disease-ADR, and drug-ADR. We developed two approaches to detect the interactions between the entities: (1) co-occurrence frequency and (2) the cosine similarity in the vector space model of word embedding. The co-occurrence frequency has been proven to be promising in heterogeneous networking mining for drug repositioning [25] and adverse drug reaction detection [26]. We used the co-occurrence frequency as a benchmark. In this work, we propose to investigate the word embedding approach because the word embedding involves the context information to determine the interactions between medical entities while the co-occurrence frequency approach only considers the simultaneous occurrence of the two corresponding entities.

(a) Co-occurrence frequency measurement – lift

In association rule mining, *lift* is a measure based on probability and reflects the division of the actual probability and theoretical probability. For instance, when measuring the strength of association rule $R \Rightarrow ADR$, lift not only takes account of $support(R \cup ADR)$ but also the the correlation between 1-itemset R and 1-itemset ADR, by calculating the ratio of the proportion of threads containing both R and ADR above those expected if R and ADR are independent of each other. There are both 1-itemset ($\{D\}$, $\{R\}$, $\{ADR\}$) and 2-itemset ($\{D, D\}$, $\{R, R\}$, $\{ADR, ADR\}$, $\{D, R\}$, $\{D, ADR\}$, $\{R, ADR\}$) involved in our calculation. The goal is to mine and evaluate the associations presented in 2-itemset. For a direct link $A_1 \leftrightarrow A_2$, the equation for calculating $lift\ (n_i, n_j)$ ($n_i \in A_1, n_j \in A_2$) is:

$$lift(n_i, n_j) = \frac{support(n_i, n_j)}{support(n_i) \times support(n_j)}$$

$$support\ (n_i) = \frac{count(n_i)}{total\ threads}$$

$$support\ (n_i, n_j) = \frac{count(n_i \cup n_j)}{total\ threads}$$

in which, $count(n_i)$ is the number of threads that contain target n_i; $count(n_i \cup n_j)$ is the number of threads that contain both n_i and n_j; *total threads* denote the total number of threads.

(b) Vector similarity in word embeddings

The basic idea of word embedding is to involve the contextual information during the learning of word vectors and to represent words with low-dimensional vectors (dimensions usually between 50 and 1000). The computation process can be summarized as: assign a random vector for each word in the vocabulary; traverse the text corpus step by step, and at each step, observe the target word and its context and update the word's and the content words' vectors to make them close in the vector space, while update other vectors to make them less close to the target word. After updating the word vectors iteratively, the vectors become meaningful and similar vectors yield to similar words. Moreover, cosine similarity is usually used to measure the similarity between two words in the vocabulary by measuring the angle between two word vectors. The cosine similarity between two nodes a and b is calculated by the equation:

$$sim(a, b) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} = \frac{\sum_{i=1}^{d} \boldsymbol{a}_i \boldsymbol{b}_i}{\sqrt{\sum_{i=1}^{d} \boldsymbol{a}_i{}^2} \sqrt{\sum_{i=1}^{d} \boldsymbol{b}_i{}^2}}$$

in which, $\boldsymbol{a}$ and $\boldsymbol{b}$ denote the vectors of word a and b, d is the dimension of the vector.

B. Heterogeneous network

1) Heterogeneous network construction

A heterogeneous network is defined as a graph that consists of at least two types of nodes or edges [27]. In real world, most networks are actually heterogeneous networks rather than homogeneous networks that address the within nodes and edges as the same type. Let $N = \{n_1, n_2, \ldots, n_k\}$ be a set of nodes and $L = \{l_1, l_2, \ldots, l_m\}$ be a set of edges, then $G = (N, L)$ denotes the graph. In the graph G, each node $n_i \in N$ belongs to a particular type from γ; each edge $l_i \in L$ belongs to a particular type from τ, and $|\gamma| > 1$ or $|\tau| > 1$. Then $M_G = (\gamma, \tau)$ denotes the node types γ and edge types τ in graph G.

By involving the medical entities we identified from the corpus, we constructed a heterogeneous network that contains three types of nodes (disease(D), drug(R), ADR) and six types of links (R-R, D-D, ADR-ADR, R-D, R-ADR, D-ADR), as shown in Fig. 1. That is, $\gamma = \{D, R, ADR\}$, and $\tau = \{L_{D-D}, L_{R-R}, L_{ADR-ADR}, L_{D-R}, L_{D-ADR}, L_{R-ADR}\}$. In this network, the interaction between two nodes, $w(n_i, n_j)$ were measured by either their co-occurrence in the same thread or the cosine similarity of their embedding vectors.

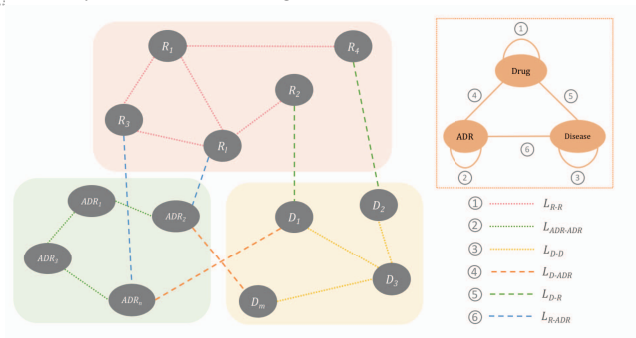

Fig. 1. Structure of heterogeneous healthcare network

2) Meta Path

A meta path is a path defined on the network schema in the form of $A_1 \xrightarrow{L_1} A_2 \xrightarrow{L_2} \dots \xrightarrow{L_l} A_{l+1}$, which composes the relations between nodes in the heterogeneous network. Meta path-based approaches could describe the structure of the paths that derived from the meta paths and the meta structure of the network. In order to infer all the possible and reliable associations between diseases and drugs, we defined the topology between them using meta paths by limiting the length within three. Only involving D and R, we determined seven meta paths, as shown in Table 1; by adding ADR, we determined extra six meta paths, as shown in Table 2.

Table 1. Meta paths via *D* and *R* only

	Length	Meta Path	Structure
1	1	*D-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_j$
2	2	*D-D-R*	$d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-R}}{\longleftrightarrow} r_j$
3	2	*D-R-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{R-R}}{\longleftrightarrow} r_j$
4	3	*D-D-D-R*	$d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-D}}{\longleftrightarrow} d_m \overset{L_{D-R}}{\longleftrightarrow} r_j$
5	3	*D-D-R-R*	$d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-R}}{\longleftrightarrow} r_m \overset{L_{R-R}}{\longleftrightarrow} r_j$
6	3	*D-R-D-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{D-R}}{\longleftrightarrow} d_m \overset{L_{D-R}}{\longleftrightarrow} r_j$
7	3	*D-R-R-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{R-R}}{\longleftrightarrow} r_m \overset{L_{R-R}}{\longleftrightarrow} r_j$

Table 2. Meta paths via *D*, *R*, and *ADR*

	Length	Meta Path	Structure
8	2	*D-ADR-R*	$d_i \overset{L_{D-ADR}}{\longleftrightarrow} adr_k \overset{L_{R-ADR}}{\longleftrightarrow} r_j$
9	3	*D-D-ADR-R*	$d_i \overset{L_{D-D}}{\longleftrightarrow} d_k \overset{L_{D-ADR}}{\longleftrightarrow} adr_m \overset{L_{R-ADR}}{\longleftrightarrow} r_j$
10	3	*D-R-ADR-R*	$d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{R-ADR}}{\longleftrightarrow} adr_m \overset{L_{R-ADR}}{\longleftrightarrow} r_j$
11	3	*D-ADR-D-R*	$d_i \overset{L_{D-ADR}}{\longleftrightarrow} adr_k \overset{L_{D-ADR}}{\longleftrightarrow} d_m \overset{L_{D-R}}{\longleftrightarrow} r_j$
12	3	*D-ADR-R-R*	$d_i \overset{L_{D-ADR}}{\longleftrightarrow} adr_k \overset{L_{R-ADR}}{\longleftrightarrow} r_m \overset{L_{R-R}}{\longleftrightarrow} r_j$
13	3	*D-ADR-ADR-R*	$d_i \overset{L_{D-ADR}}{\longleftrightarrow} adr_k \overset{L_{ADR-ADR}}{\longleftrightarrow} adr_m \overset{L_{R-ADR}}{\longleftrightarrow} r_j$

3) Meta-path-based Indicators

Given the topological features determined by meta paths, the associations between D and R can be measured based on the commonly used indicator – Path Count (PC), which counts the number of path instances between two nodes under a given meta path. More than just counting the number of paths, here we incorporated more network information by taking into account of the weights of edges and proposed two indicators to describe the associations between two nodes d_i and r_j ($d_i \in D, r_j \in R$): given a meta path P, (1) Path-Count-Lift (PCL) measures the weight of meta paths with *lift* and sums up the weights of all the p ($p \in P$) that associate associates d_i with r_j; (2) Path-Count-Embedding (PCE) measures the weight of meta paths with *sim* and sums up the weights of all the p ($p \in P$) that associate associates d_i with r_j.

Given a meta path P in the form of $A_1 \xrightarrow{L_1} A_2 \xrightarrow{L_2} \dots \xrightarrow{L_l} A_{l+1}$, $L_P(n_1, n_l)$ ($n_1 \in A_1, \dots, n_l \in A_l$), Path-Count-Lift (PCL) is calculated by:

$$PCL_P(n_1, n_l) = \sum_{p \in P} lift(n_1, n_2) \times lift(n_2, n_3) \times \dots \times lift(n_{l-1}, n_l), (n_1 \overset{p}{\leftrightarrow} n_l = 1)$$

Path-Count-Embedding (PCE) is calculated by:

$$PCE_P(n_1, n_l) = \sum_{p \in P} sim(n_1, n_2) \times sim(n_2, n_3) \times \dots \times sim(n_{l-1}, n_l), (n_1 \overset{p}{\leftrightarrow} n_l = 1)$$

For example, if given the meta path *D-R-R*:

$$PCL_{D-R-R}(d_i, r_j) = \sum_{p \in P} lift(d_i, r_k) \times lift(r_k, r_j), (d_i \overset{p}{\leftrightarrow} r_j = 1)$$

$$PCE_{D-R-R}(d_i, r_j) = \sum_{p \in P} sim(d_i, r_k) \times sim(r_k, r_j), (d_i \overset{p}{\leftrightarrow} r_j = 1)$$

in which, $d_i \overset{p}{\leftrightarrow} r_j = 1$ denotes there exists a meta path p that associates d_i and r_j via $d_i \overset{L_{D-R}}{\longleftrightarrow} r_k \overset{L_{R-R}}{\longleftrightarrow} r_j$. In addition, we also calculated Path Count (PC) as the baseline indicator:

$$PC_P(d_i, r_j) = \sum_{p \in P} 1, (d_i \overset{p}{\leftrightarrow} r_j = 1)$$

C. Classification

We trained a binary classifier to recognize the known disease-drug associations from all the possible pairs. The inputs of the classifier were derived from the meta-path-based indicators and the outputs were either positive or negative, representing whether the predicted disease-drug pair could be a known usage or not. In specific, for each disease-drug pair we built a set of features that are used to classify if this pair is possibly to be a known association. As we defined two meta-path-based indicators (i.e. *PCL* and *PCE*), we used them respectively and their combinations to be features to train the classifiers, as well as the baseline indicator. In this study, we utilized the popular machine learning method Random Forest as the classification algorithm, which has been proved to be effective and powerful for using network-based features to predict biomedical relationships.

D. Detecting off-label drug uses

Resources from medication providers such as clinical notes and EHRs provide the opportunity to detect off-label uses in an automated and scalable way, meanwhile, the large volume of contents generated by medication receivers such as social media and OHC data contribute another valuable resource to detect off-label uses systematically. The hypothesis here is that if the features extracted from user generated data enable us to recognize the known drug-disease usages effectively, the other drug-disease pairs that show similar features with the known usages have a high possibility to be off-label practices, that is, the negative associations that are falsely classified as positive are potential to be the off-label drug-disease usages. Therefore, the task here is to identify the false positive (FP) predictions in the confusion matrix of classification.

IV. EXPERIMENT & RESULTS

In this work, we collected the user generated data from MedHelp (www.medhelp.org), a pioneer in OHCs and owning 176 health communities on the site. We retrieved all the posts and comments within the most popular 50 disease communities by operating an automatic web crawler, returning more than 70,000 posts and 319,000 comments. Then we detected 50 diseases, 1,297 drugs, and 185 ADRs from the text with the lexicon-based approach. On the corpus, we trained the word embeddings using *word2vec*, the state-of-the-art word embedding model based on neural networks, and

obtained an embedding model containing 356,776 words and their corresponding 200-dimensional vectors. With *lift* and *sim* denoting the interactions between medical entities, we constructed two heterogeneous networks and computed the defined meta-path-based indicators.

Then we created a gold standard dataset to implement the supervised classification model, in which, the positive pairs or known drug-disease usages were extracted from PharmGKB and DrugBank, and the negative pairs were generated randomly from the unobserved associations between 50 diseases and 1,297 drugs. As a result, the dataset contained 2,087 known drug-disease usages and 28,000 negative pairs. Considering there were much more negative instances than the positive in the dataset, we operated undersampling to deal with the imbalanced classification problem. Firstly, we divided the whole gold standard dataset into training (65%) and test (35%) sets; secondly, we randomly split the negative instances into 10 chunks, each chunk and the positive instances in training set composed a sub-training set. Thus, the classifiers were trained on each sub-training set and evaluated on the hold-out test set, and the overall classification performance were represented by the average of 10 classifiers trained on 10 sub-training sets.

A. Classification results

The classification performance was evaluated by Precision, Recall, and F1-score. We trained and tested the Random Forest classifier with *scikit-learn* tool in Python package (Pedregosa et al., 2011), using four groups of features described in Section 3.3. The evaluation results were shown in Table 3:

Table 3 Evaluation results of different classification models

Feature	Precision	Recall	F1-score
PC	0.763	0.745	0.754
PCL	0.795	0.969	0.870
PCE	0.820	**0.986**	0.895
PCL+PCE	**0.908**	0.973	**0.939**

*Bold indicates the highest score in the column

The results showed that using meta-path-based topological features to classify drug-disease associations was effective and obtained a quite acceptable performance with the lowest F1-score getting 0.754 in the hold-out test set. Compared with the other indicators, PC performed the worst when used as classification features on all the three evaluation measures. When incorporating information about the weights of edges and paths by using PCL and PCE, F1-scores were improved by at least 15% (>0.87), which indicated a distinct overall improvement of the classification model, meanwhile, Recalls were increased by 30% (>0.96), which means that among all the known drug-disease usages we classified over 96% of them correctly. When comparing PCL and PCE, the overall performance of using PCE was lightly better than that of PCL according to all three measures, which might suggest that the cosine similarity of word embedding vectors between nodes described their relationship better than lift. Additionally, when combining PCL and PCE features to feed in the classifier, it achieved the best performance according to F1-score (0.939), with Recall of 0.973 indicating that among all the known drug-disease usages 97.3% of them were classified correctly, and Precision of 0.908 indicating that 90.8% of the predicted "positive" drug-disease pairs were indeed known usages and the other 9.2% that were falsely classified as "positive" from the randomly generated negative pairs might be the off-label pairs we were searching for.

Since using PCL and PCE together for classification contributed the best performance, we then applied the best trained classification model to the whole balanced dataset built by oversampling the minority class, to find all the possible off-label predictions. In result (shown in Table 4), we found 2,125 false-positive instances that have the potential candidates of off-label drug-disease associations.

Table 4: Classification results of using PCL+PCE to classify the whole dataset

		Predicted	
		P	N
Actual	P	20243	627
	N	2125	25875

B. Validation of off-label use candidates

We examined the positive evidence for the detected 2125 potential off-label drug uses in PubMed and FAERS. PubMed is a publicly available repository managed by National Center for Biotechnology Information (NCBI) and covers the titles and abstracts of more than 26 million biomedical publications, which can be accessed with the provided Entrez Programming Utilities (E-utilities). FAERS is the most important spontaneous reporting system as well as the primary data source for the study and identification of ADRs in United States. In PubMed, we found 821 of the novel predictions that have at least ten articles that both the drug and the disease were mentioned in the abstract; in FAERS, we only found 35 of the novel predictions that have at least ten reports co-annotating with both the drug and the disease. The reason why there was so less evidence in FAERS might be that: first, FAERS is designed for reporting adverse drug effects rather than off-label uses, therefore, if there is no ADR involved, people have no intension to report their situations to the system; second, people report ADRs spontaneously and voluntarily, which leads to a surprisingly low reporting rate because of the nature of passiveness, with a median of 6%; third, it usually takes FDA a long time to complete the whole process of collecting reports, investigating cases and releasing alerts, which limits the manner of timely for information.

V. CONCLUSION

In clinical practice, off-label drug uses are very common and inevitable to some extent. In addition, the stakeholders such as drug companies, healthcare professionals, and researchers all have the need to get information about off-label drug uses timely. Therefore, it raises the demand for a systematic way to detect off-label drug uses. The data coming from healthcare providers such as clinical notes and EHRs provide the resource of detecting off-label uses and have been utilized in previous studies, meanwhile, the large volumes of data generated by healthcare receivers also offer the great

opportunity to detect off-label uses in an automated and scalable way. In this work, we proposed a systematic method to detect off-label drug uses from health consumer contributed data based on meta-path-based heterogeneous network mining and binary classification. With data collected from a popular OHC-MedHelp, we extracted the medical entities (diseases, drugs, and ADRs) with lexicon-based approaches and measured the interactions between them by using association rule mining and word embedding. Then we constructed a heterogeneous healthcare network with those entities as nodes and interactions as edges, in which, we determined 13 meta paths between diseases and drugs and defined two meta-path-based indicators to describe the disease-drug associations: Path-Count-Lift (PCL) and Path-Count-Embedding(PCE). Then we utilized these features as inputs for the Random Forest classifier to recognize the known drug-disease associations from all the possible pairs. Using Path Count, PCL, PCE, and PCL+PCE to be the features respectively, we implemented four classification experiments, and the model built on PCL+PCE obtained the best performance with F1-score reaching 0.939. The results indicated that meta-path-based network features can be used for developing effective supervised classifiers of identifying known drug-disease associations, especially when incorporating text features with word embedding models. Furtherly, the other drug-disease pairs that show similar features with those known pairs are potential to be the off-label practices, that is, the false-positive predictions are potential to be the off-label drug-disease usages. Based on such hypothesis, we identified 2,125 potential candidates of off-label drug uses from classification results, and then examined their potential using PubMed abstracts and FAERS reports.

VI. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under the Grant IIS-1650531 and DIBBs-1443019.

REFERENCES

[1]. Stafford, R. S. (2008). Regulating off-label drug use—rethinking the role of the FDA. *New England Journal of Medicine*, *358*(14), 1427-1429.

[2]. Wittich, C. M., Burkle, C. M., & Lanier, W. L. (2012, October). Ten common questions (and their answers) about off-label drug use. In *Mayo Clinic Proceedings* (Vol. 87, No. 10, pp. 982-990). Elsevier.

[3]. Jung, K., LePendu, P., Chen, W. S., Iyer, S. V., Readhead, B., Dudley, J. T., & Shah, N. H. (2014). Automated detection of off-label drug use. PloS one, 9(2), e89324.

[4]. Conroy, S., Choonara, I., Impicciatore, P., Mohn, A., Arnell, H., Rane, A., ... & Rocchi, F. (2000). Survey of unlicensed and off label drug use in paediatric wards in European countries. Bmj, 320(7227), 79-82.

[5]. Leslie, D. L., & Rosenheck, R. (2012). Off-label use of antipsychotic medications in Medicaid. The American journal of managed care, 18(3), e109-17.

[6]. Alexander, G. C., Gallagher, S. A., Mascola, A., Moloney, R. M., & Stafford, R. S. (2011). Increasing off-label use of antipsychotic medications in the United States, 1995–2008. Pharmacoepidemiology and drug safety, 20(2), 177-184.

[7]. Turner, S. E. A. N., Nunn, A. J., Fielding, K., & Choonara, I. M. T. I. (1999). Adverse drug reactions to unlicensed and off-label drugs on paediatric wards: a prospective study. Acta Paediatrica, 88(9), 965-968.

[8]. Neubert, A., Dormann, H., Weiss, J., Egger, T., Criegee-Rieck, M., Rascher, W., ... & Hinz, B. (2004). The impact of unlicensed and off-label drug use on adverse drug reactions in paediatric patients. Drug safety, 27(13), 1059-1067.

[9]. Horen, B., Montastruc, J. L., & Lapeyre-Mestre, M. (2002). Adverse drug reactions and off-label drug use in paediatric outpatients. British journal of clinical pharmacology, 54(6), 665-670.

[10]. Eguale, T., Buckeridge, D. L., Verma, A., Winslade, N. E., Benedetti, A., Hanley, J. A., & Tamblyn, R. (2016). Association of off-label drug use and adverse drug events in an adult population. JAMA internal medicine, 176(1), 55-63.

[11]. Jung, K., LePendu, P., & Shah, N. (2013). Automated detection of systematic off-label drug use in free text of electronic medical records. AMIA Summits on Translational Science Proceedings, 2013, 94.

[12]. Mesgarpour, B., Müller, M., & Herkner, H. (2012). Search strategies to identify reports on "off-label" drug use in EMBASE. BMC medical research methodology, 12(1), 190.

[13]. Xu, R., & Wang, Q. (2013). Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC bioinformatics, 14(1), 181.

[14]. Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular systems biology, 7(1), 496.

[15]. Wang, W., Yang, S., Zhang, X., & Li, J. (2014). Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics, 30(20), 2923-2930.

[16]. Martínez, V., Navarro, C., Cano, C., Fajardo, W., & Blanco, A. (2015). DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. Artificial intelligence in medicine, 63(1), 41-49.

[17]. Huang, Y. F., Yeh, H. Y., & Soo, V. W. (2013). Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. BMC medical genomics, 6(3), S4.

[18]. Chen, X., Liu, M. X., & Yan, G. Y. (2012). Drug–target interaction prediction by random walk on the heterogeneous network. Molecular BioSystems, 8(7), 1970-1978.

[19]. Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., & Collier, N. (2004, August). Introduction to the bio-entity recognition task at JNLPBA. In Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (pp. 70-75). Association for Computational Linguistics.

[20]. Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics*, *6*(1), 1.

[21]. Wang, J., Zhang, J., An, Y., Lin, H., Yang, Z., Zhang, Y., & Sun, Y. (2015, November). Biomedical event trigger detection by dependency-based word embedding. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on* (pp. 429-432). IEEE.

[22]. Li, C., Song, R., Liakata, M., Vlachos, A., Seneff, S., & Zhang, X. (2015). Using word embedding for bio-event extraction. *ACL-IJCNLP 2015*, 121.

[23]. Jiang, L., & Yang, C. C. (2015, March). Expanding Consumer Health Vocabularies by Learning Consumer Health Expressions from Online Health Social Media. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 314-320). Springer International Publishing.

[24]. Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, *13*(1), 24-29.

[25]. Zhao, M., & Yang, C. C. (2016, October). Mining Online Heterogeneous Healthcare Networks for Drug Repositioning. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on* (pp. 106-112). IEEE.

[26]. Yang, H., & Yang, C. C. (2016, October). Discovering Drug-Drug Interactions and Associated Adverse Drug Reactions with Triad Prediction in Heterogeneous Healthcare Networks. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on* (pp. 244-254). IEEE.

[27]. Han, J., Sun, Y., Yan, X., & Yu, P. S. (2010, July). Mining heterogeneous information networks. In Tutorial at the 2010 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'10), Washington, DC.