Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks

Kathy Lee[1,2], Ashequl Qadir[1], Sadid A. Hasan[1], Vivek Datla[1],
Aaditya Prakash[1,3], Joey Liu[1], Oladimeji Farri[1]
[1]Artificial Intelligence Laboratory, Philips Research North America, Cambridge, MA, USA
{kathy.lee_1, ashequl.qadir, sadid.hasan, vivek.datla, aaditya.prakash, joey.liu, dimeji.farri}@philips.com
[2]Northwestern University, kathy.lee@eecs.northwestern.edu
[3]Brandeis University, aprakash@brandeis.edu

ABSTRACT

Current Adverse Drug Events (ADE) surveillance systems are often associated with a sizable time lag before such events are published. Online social media such as Twitter could describe adverse drug events in real-time, prior to official reporting. Deep learning has significantly improved text classification performance in recent years and can potentially enhance ADE classification in tweets. However, these models typically require large corpora with human expert-derived labels, and such resources are very expensive to generate and are hardly available. Semi-supervised deep learning models, which offer a plausible alternative to fully supervised models, involve the use of a small set of labeled data and a relatively larger collection of unlabeled data for training. Traditionally, these models are trained on labeled and unlabeled data from similar topics or domains. In reality, millions of tweets generated daily often focus on disparate topics, and this could present a challenge for building deep learning models for ADE classification with random Twitter stream as unlabeled training data. In this work, we build several semi-supervised convolutional neural network (CNN) models for ADE classification in tweets, specifically leveraging different types of unlabeled data in developing the models to address the problem. We demonstrate that, with the selective use of a variety of unlabeled data, our semi-supervised CNN models outperform a strong state-of-the-art supervised classification model by +9.9% F1-score. We evaluated our models on the Twitter data set used in the PSB 2016 Social Media Shared Task. Our results present the new state-of-the-art for this data set.

Keywords

Adverse drug events; pharmacovigilance; text classification; social media; semi-supervised convolutional neural networks; healthcare

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052671

1. INTRODUCTION

Adverse Drug Events (ADE) refer to experiencing an injury that may occur as a result of a drug use.[1,2] Monitoring and detection of such events (also called Pharmacovigilance) is necessary to minimize potential health risks of patients by issuing warnings with the relevant pharmaceutical products. Following pharmaceutical development, drugs are typically approved for use after going through clinical trials in limited settings. It is often impossible to uncover all adverse effects during these clinical trials. To address this issue, pharmaceutical and regulatory organizations require post-market surveillance to capture previously undiscovered side effects. However, current post-market ADE surveillance systems are associated with under-reporting and significant time delays in data processing, resulting in high incidence of unidentified adverse events related to drug use [35].

In the past decade, the rise of social media platforms (e.g., Twitter) has revolutionized our online communication and networking, and has been used for real-time information retrieval and trends tracking, including digital disease surveillance [26, 24, 25]. Twitter is one of the popular forms of social media and a potential resource for detecting ADEs in real-time. However, there are several challenges in detecting ADEs in Twitter posts. For example, (1) sparsity of ADE tweets in real-world Twitter stream,[3] (2) health conditions described in colloquial language, (3) mentions of side effect-like conditions (we will refer to these phrases as "health conditions" in this paper) and drugs in the same tweet without necessarily representing an ADE.

Table 1 demonstrates some of these challenges in real tweets. *'Face burning up'* would be rare to find in a standard pharmaceutical dictionary, while it is a colloquial term for 'facial burning sensation'. Despite the mention of health conditions (e.g., *'My face is on fire'*, *'I'm burning up'*) as well as a drug name (e.g., Tylenol), the second tweet does

[1] `http://www.va.gov/MS/Professionals/medications/Adverse_Drug_Reaction_FAQ.pdf`

[2] Note that there are subtle differences among ADR (Adverse Drug Reaction), ADE (Adverse Drug Event) which includes ADR and effects from overdose or discontinuation of medication, etc., and Side Effects (unintended reaction from medication). From a practical point of view, distinguishing semantic differences among these is a much harder challenge. We use ADE as an umbrella term to refer to all of the above.

[3] Only 11.4% of tweets that mention a drug are ADE tweets [37]. We estimated that in a random tweets sample, common drug names (from Walgreens) appeared in only 0.013% tweets.

Table 1: Example of ADE and non-ADE Tweets

Class	Tweet
ADE	Oh yay, **Niaspan** reaction. *Face burning up.*
Non-ADE	*My face is on fire* and **Tylenol** isn't helping. *I'm burning up.* Fingers crossed I'm not getting sick.

not represent an ADE. This is because the phrase 'Tylenol isn't helping' implicitly tells that the user took Tylenol to relieve his/her 'facial burning sentation' (an *indication* event instead of ADE). Prior research in ADE detection is typically focused on solving: 1) binary classification of tweets or sentences that mention an ADE, and/or 2) extraction of ADE phrases (e.g., *I'm burning up*). Our work addresses the first task.

Recently, deep learning models have been successful [18, 23, 38] and gained popularity because of both superior performance and for not requiring domain/problem-specific feature engineering unlike traditional machine learning algorithms. However, deep learning models typically rely on some large annotated corpora for good performance. Given the sparsity of ADE tweets in real-world Twitter stream, such resources are expensive to generate for ADE tweet classification. Semi-supervised approaches to training deep learning models offer a plausible alternative to fully supervised models. They involve the use of a small set of labeled data and relatively larger collections of unlabeled data for training. But for ADE tweet classification, the impact of unlabeled data is unknown as random microblogs will have many disparate topics that may not be relevant to ADE.

In this work, we use a semi-supervised Convolutional Neural Network (CNN)-based architecture [11] to build several semi-supervised CNN models for ADE classification in tweets, specifically leveraging different types of unlabeled data. In particular, we present experiments with models that use a large collection of random tweets, tweets with drug names, health conditions, sentences from scientific articles in the medical literature and Wikipedia, simulated health-related sentences created from lexicons, and combinations of these data types. We demonstrate that, with the selective use of a variety of unlabeled data, our classification models outperform strong supervised classification models (that use the same amount of labeled data as our models) from previous work, improving ADE classification by +9.9% F1-score (over the previous best model) on a Twitter data set.

In summary, our key research contributions are:

- We explore semi-supervised convolutional neural network (CNN) for automatic classification of ADE tweets.
- We investigate the use of various types of text data that are problem specific, or from health-related domains. We present experimental results by using these data individually and in combination as the unlabeled data for the semi-supervised CNN models.
- We demonstrate that, by training models with selective use of unlabeled data, and combining problem-specific tweets and health-related sentences instead of using a large collection of general random tweets, ADE tweet classification performance can be improved substantially.
- Our best results outperform the results of a state-of-the-art supervised ADE classification model [37] by +14.58% precision, +6.02% recall, and +9.9% F1-score when evaluated on the same Twitter data set, which was also used in the PSB 2016 Social Media Mining Shared Task[4] and in several prior works [37, 5, 1]. To the best of our knowledge, our results present the new state-of-the-art for this data set.

This paper is organized as follows. In section 2, we present related work on the adverse drug event detection in social media, CNN and semi-supervised learning. In section 3, we describe the semi-supervised CNN model for ADE classification, how we create models with various unlabeled data sets, and our experimental setup. In section 4, we present and compare performance of our models. Finally, we conclude our findings in section 5.

2. RELATED WORK

2.1 ADE Detection in Social Media

With an increasing popularity of microblog sites where people post their health-related experiences, and the release of ADE-annotated corpora on patient-reported social media data [17, 37] and medical case reports [9], research on ADE detection gained much attention in recent years. For binary ADE classification, a popular approach is to use supervised classifiers with a wide variety of features that are relevant for the problem [37, 42, 34, 32, 13, 7, 5, 33]. The lexical features used in previous studies typically include word and character n-grams, parts-of-speech tags, selective use of n-grams exploiting mutual information or term frequency-inverse document frequency; semantic features include UMLS semantic types that represent medical concepts, mentions of chemical substance and disease, WordNet synsets, adverse drug reaction lexicon; and sentiment features include phrases denoting change in sentiment, sentiment polarity cues, emotion classes, etc.

Other types of features explored for supervised classification include relational features such as co-occurrence of drugs and side effects, topical features such as topic model-based features, word embeddings-based and word cluster-based features, or language features such as, negation, text length, presence of comparative/superlative, modals, etc. In addition to traditional machine learning classification models, researchers also used ensemble classification [34, 42] or explored lexical normalization to reduce language irregularity [32]. To the best of our knowledge, ours is the first work to explore a semi-supervised deep learning classification framework for ADE detection in tweets. We compare our model's performance with a state-of-the-art supervised classification model from previous work [37].

The ADE phrase extraction task has also been studied by several researchers. Machine learning-based concept extraction systems using CRF (Conditional Random Fields) [31] or HMM (Hidden Markov Model) [35] have been implemented to extract mentions of ADEs from online user-generated data. Also, the correlation between the ADEs in tweets and the official reports by FDA Adverse Event Reporting System (FAERS) [8], building dictionary of ADE phrases using bootstrapping [1], and normalization of medical concepts describing ADEs to a controlled vocabulary in

[4] http://diego.asu.edu/psb2016/task1data.html

SNOMED-CT [20, 21] have been explored. Others have extracted named entities for medical concepts (drugs and adverse effects) from biomedical literature and microblog posts and identified whether relationships between these medical concepts represent ADEs [16, 40]. A more detailed overview of the research on this topic can be found in [19, 36].

2.2 CNN and Semi-Supervised Learning

CNN is a feed-forward neural network model that has typically been used for image processing. CNN automatically learns features, such as edges and shapes in images, that are important for the intended task (e.g., image classification) [22]. In the past few years, deep CNN models have gained attention as it has been very effective in NLP (natural xflanguage processing) tasks such as sentence classification. It has been demonstrated that a simple CNN model with pre-trained word vectors can improve performance beyond state-of-the-art algorithms on tasks like question classification [18]. Many other studies have explored the use of CNNs for text processing including tweets [6, 41, 15].

Recently, a semi-supervised CNN framework has been proposed [11] that, unlike methods that entirely rely on word embeddings, learns region embeddings (low dimensional representation of text regions) of high-level concepts of the problem. This method outperformed previous best classification results on three data sets (IMDB - movie reviews [28], Elec - Amazon reviews of electronics products, RCV1 - Reuters news articles data sets). In this work, we use this semi-supervised CNN framework for ADE classification in tweets, but leverage on different types of unlabeled data so that potential high-level concepts of the problem are sufficiently represented in the text to effectively learn the region embeddings. We particularly explore building different unlabeled data sets from general and health-related domains.

Within the semi-supervised classification framework, Long Short-Term Memory (LSTM) models have also been recently explored that rivaled or improved text classification performance over CNNs [12]. However, the improvements (when found) were typically within 1% of classification error rates in most experiments. Since our work mainly focuses on the effective learning of the region embeddings using task-relevant data sets, either of the deep learning models could be used. We leave LSTM-based semi-supervised ADE classification for future work.

3. ADE TWEET CLASSIFICATION WITH SEMI-SUPERVISED CNN

To classify tweets that indicate adverse drug events, we use a semi-supervised Convolutionational Neural Network-based (CNN) architecture (shown in Figure 1), recently proposed for text classification [11]. The method works in two phases: (1) unsupervised phrase embedding learning, and (2) integrating the learned embeddings into the supervised training that uses labeled data. Our main research contributions are in creating models in (1) with different data sets that can benefit the learning of phrase/region embeddings of the higher-level problem concepts.

3.1 Unsupervised Phrase Embeddings

The goal of unsupervised phrase embeddings learning phase is to learn region embeddings of task-relevant high-level concepts. The learning goal is to preserve the predictive structure of these text regions and learn their low dimensional vector representations. This framework is particularly suitable for our task because many of the health conditions that describe an adverse drug event are multiword phrases, and learning their vector representations can be beneficial for ADE classification. In this problem context, we envision these high-level concepts to be health conditions, drug mentions, or health concept related phrases that indicate an ADE (e.g., 'Niaspan **reaction**' in Table 1).

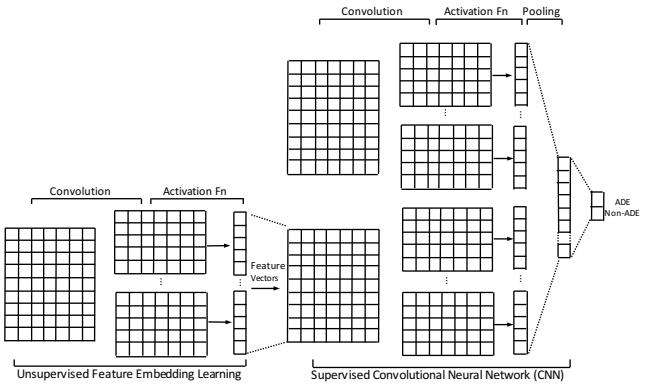

Figure 1: Semi-Supervised CNN.

Formally, given an input tweet x of length n, where $x = w_0, w_1, \ldots, w_n$ denotes a sequence of words, the neural network learns a vector representation for the i-th phrase $p_i(x)$ of size s (where $p_i(x) = w_i, w_{i+1}, \ldots, w_{i+s-1}$) by predicting its surrounding contexts. For example, when $s = 3$, for the tweet *"I took Aleve and now my stomach hurts. not good!"*, the network can learn a vector representation of the text region/phrase *'my stomach hurts'*, that allows to predict the context words/regions such as *'Aleve and now'*, *'not good'*, etc. At the time of learning, $p_i(x)$ is first converted into a vector representation. Let, $r_i(x)$ be the input region vector (bag-of-word vector) for $p_i(x)$ and $u_i(x)$ be the region embedding vector to be learned from unlabeled data for $p_i(x)$, $\boldsymbol{W}$ be the corresponding weight matrices, $\boldsymbol{b}$ be the bias vector, and σ be the component-wise non-linear activation function; then learning for $u_i(x)$ can be formulated as follows:

$$u_i(x) = \sigma(\boldsymbol{W} \cdot r_i(x) + \boldsymbol{b}) \quad (1)$$

For more details about the formulation, see [11]. The learned region embeddings are then integrated into the supervised CNN.

3.2 Data Selection for Unsupervised Pre-training

Intuitively, for classifying ADE tweets, our model may benefit from learning embeddings of text regions that represent drugs names, health conditions, and phrases that describe an ADE-relationship between two or more concepts. So, we need to selectively focus on providing relevant data to the model to allow learning embeddings for text regions representing these high-level concepts. To this end, we build several semi-supervised CNN models with varying types of unlabeled data (including the one with a massive collection of random tweets for comparison). In the following sections, we present these different models built for the unsupervised phrase embedding learning part of the semi-supervised CNN.

3.2.1 Model-1 T-Random

Our first model uses a massive collection of random tweets for learning the phrase embeddings. We build this model mainly for comparison purposes, as we hypothesize that unlabeled data of such general nature is not sufficient for learning the problem specific region embeddings effectively.

First we create a corpus of random tweets using the Twitter streaming API,[5] which provides approximately 1% of all publicly available tweets. The tweets were collected over different time periods in 2015 and 2016. As preprocessing steps, we tokenize the text, normalize to lowercase, remove duplicate tweets and tweets shorter than five words.[6] Specifically for Twitter data, we also remove non-English tweets and retweets. We replace hyperlinks and Twitter screen names with special tokens: 'URL' and 'USER'. The final tweets data set (*T-Random*) consists of about 43.5 million tweets. We use these tweets as the unlabeled data to train our *Model T-Random*.

3.2.2 Model-2 Sent-Health

Our next model (*Model Sent-Health*) uses a corpus of health-related texts containing a collection of sentences from the following sources in medical domain.

Medical Concept to Lay Term Dictionaries: We use two medical to lay terms dictionaries to create a collection of sentences.[7,8] These dictionaries contain professional medical terms and their definitions described in lay language. For example, the medical term 'anesthesia' is defined in lay language as 'loss of sensation or feeling', the term 'cephalalgia' as 'headache', and the term 'dyspnea' as 'hard to breathe' or 'short of breath'. From these dictionaries, we generate sentences (e.g., 'Anesthesia refers to loss of sensation or feeling', 'cephalalgia means headache') by combining a term and its definition with a connecting phrase randomly chosen from a small preselected set (e.g., stands for, refers to, indicates, means, etc.). We create a total of 1,556 sentences from these sources.

Medical Concept Terms for Social Media: Medical concepts are often described differently in online social media. Conceptually, medical to lay terms dictionaries are created for easier understanding of the medical terms, but these lay terms are not necessarily represented in social media, especially when users attempt to describe their personal experiences. We use all sentences from a corpus that contains annotations for medical concepts, as written in social media [17]. We also take a medical term normalization data set [27] that contains a mapping between a health condition and the corresponding concept in standard ontology such as SNOMED CT[9] that contains clinical terms used by medical professionals (e.g., 'head spinning a little' → 'dizziness', 'my foot feel worse' → 'foot pain', 'my eyes won't stay open' → 'drowsiness'). We create sentences for these mappings similar to sentences from the lay term dictionaries. Combining sentences created using these two dictionaries, we obtain a total of 8,432 sentences.

Biomedical Literature: We collect 301,790 sentences from all wikipedia pages that are under the category of clinical medicine.[10] We also collect 4,271 sentences from PubMed articles that mention ADEs in the ADE benchmark corpus [9].

UMLS Medical Concept Definitions: We extract a total of 167,550 sentences that define medical terms in the UMLS Metathesaurus [3], a large biomedical thesaurus consisting of millions of medical concepts and used by professionals for patient care and public health.

We combine all of the above mentioned sentences to build the unlabeled data set, *Sent-Health* (483,599 sentences), for training the *Model Sent-Health*.

3.2.3 Model-3 T-Drug

Our third model (*Model T-Drug*) is trained using tweets with drug names as the unlabeled data. To build this corpus, we first create a dictionary of commonly used drug names, with a goal to acquire tweets that mention one of these drug names. We collected 2,566 drug names from Walgreens pharmacy website.[11] However, we found that many drug names were polysemous with other dominant senses, posing a challenge for using them to collect tweets. Ideally, one could apply a word sense disambiguation (WSD) system to determine when a drug name is actually used as a drug, which we do not explore in this work.

We use a simple heuristic to filter out any drug name that also exists in WordNet [30] but is not in the hyponym tree of 'drug', 'agent' or 'substance'. This heuristic particularly does not work when a polysemous drug name is not a noun in WordNet or does not appear in WordNet. So, from the resulting list, we manually removed a small number of additional drug names that are common words (e.g., Advanced, Compete), tend to be people's names (e.g., Yasmin), and drug names that are shorter than 4 characters (e.g., FEM). We also filtered out any drug name that never appeared in our random tweets corpus (*T-Random*). The final drugs dictionary consists of 380 drug names. We build this dictionary using drug names from drug stores like Walgreens because many formal drug names, such as the ones from the UMLS Metathesaurus [39], do not appear in Twitter as often, likely because drug brand names (e.g. advil, tylenol, benadryl) are more common for people to use than the lengthier and complex generic drug names.

Using the drug dictionary, we first extract all tweets with drug name mentions from *T-Random*. We also collect additional tweets that mention these drugs using the Twitter Search API.[12] We combine tweets from both collections. The final data set (*T-Drug*) contains 157,136 tweets that mention a drug name. These tweets are then used as the unlabeled data to train *Model T-Drug*.

3.2.4 Model-4 T-Health-Condition

For our fourth model, we use tweets that mention a health condition, as the unlabeled data. Similar to the process for *T-drug*, we first create a dictionary of health condition phrases. To build the dictionary, we combine all medical concepts in the Medical Dictionary for Regulatory Activi-

[5] https://dev.twitter.com/streaming/public

[6] These preprocessing steps are performed on all of our data sets.

[7] http://gsr.lau.edu.lb/irb/forms/medical_lay_terms.pdf

[8] https://depts.washington.edu/respcare/public/info/Plain_Language_Thesaurus_for_Health_Communications.pdf

[9] https://www.nlm.nih.gov/healthit/snomedct/index.html

[10] https://en.wikipedia.org/wiki/Category:Clinical_medicine

[11] https://www.walgreens.com/pharmacy/marketing/library/finddrug/druginfobrowseresults.jsp

[12] https://dev.twitter.com/rest/public/search

ties (MedDRA) [4], *Adverse Drug Reaction* annotations in PubMed articles [9], annotated *Adverse effect* phrases in social media and medical forums posts [17], and phrases from the medical term normalization data set [27].

From this collection, we remove phrases not found in our random tweets corpus (*T-Random*). The final dictionary consists of 7,193 health conditions. Note that the dictionary is noisy as it contains several terms that are not medical conditions (e.g., theft, delivery, poverty).[13] We consider the process of removing the noise in this data set non-trivial, and a topic for future work.

We then retrieve all tweets from *T-Random* that have mentions of these health condition phrases, and additional tweets with mentions of these phrases from Twitter via the search API. The combined collection (*T-Health-Condition*) contains a total of 3,524,350 tweets. We then use these tweets as the unlabeled data for building our *Model T-Health-Condition*.

3.2.5 Model-5 T-Health-Condition*

One of the major challenges in ADE classification is that a self-reported health condition in social media posts is often hard to recognize. The same health condition can be described in many different ways and these phrases rarely appear in a standard medical dictionary. For example, the medical condition 'insomnia' can be described as 'can't sleep', 'wide awake', 'up all night', 'sleeplessness', 'no sleep' and so on. Although we created our dictionary of health conditions combining phrases from various sources, we acknowledge that the dictionary is not exhaustive. So we explored expanding our dictionary using a neural phrase embedding model[14] which is based on the skip-gram model architecture [29, 10].

We first train the model with a large text corpus created by combining random tweets (described in Section 3.2.1) and health-related texts (described in Section 3.2.2). For each phrase in our dictionary of health conditions, we query the model for the most similar phrases and corresponding cosine similarity scores in the vector space. All output with similarity score 0.7 or higher (to ensure strong relevance) are added to our dictionary of health conditions to create an expanded health conditions dictionary.

Some examples of the similar phrases learned are: 'fatal' linked to *life-threatening, severe, serious*, and *acute*. Phrases learned using the neural embeddings represent semantically related phrases that include, but not restricted to, synonyms. The expanded dictionary also contains noise, some propagated from the noise in the initial health conditions dictionary, and others from the neural phrase embedding model for retrieving phrases that are not similar. Note that this step is different from the unsupervised feature/phrase embedding learning phase of semi-supervised CNN described in section 3.1, and outside of the semi-supervised classification framework.

The final 3,042 similar phrases are combined with the original health conditions dictionary to create the expanded dictionary. As before, we retrieve all tweets from the *T-Random* corpus, and also search in Twitter using the search API for tweets that mention these phrases from the expanded dictionary. The combined collection (*T-Health-Condition**) contains a total of 7,039,386 tweets. We train the semi-supervised CNN framework with these tweets for our *Model T-Health-Condition**.

3.2.6 Model-6 T-Drug-Condition-Sent-Health

For this model, we combine the tweets from *T-Drug*, *T-Health-Condition*, and the sentences from *Sent-Health*. The combined collection is used as the unlabeled data to train the *Model T-Drug-Condition-Sent-Health*.

3.2.7 Model-7 T-Drug-Condition*-Sent-Health

In our last model, we combine multiple data sets similar to training *Model T-Drug-Condition-Sent-Health*, but replace the tweets from *T-Health-Condition* corpus with those from the *T-Health-Condition**. We then use the combined collection of tweets and sentences as the unlabeled data for building our *Model T-Drug-Condition*-Sent-Health*.[15]

3.3 Supervised Convolutional Neural Network

In the second phase, a convolutional neural network is trained with both the phrase embeddings learned for our different models, and the annotated ADE data (tweets and their labels) to generate an *ADE classifier*.

Let $r_i(x)$ be the input region vector (bag-of-word vector) for the i-th phrase $p_i(x)$ of tweet x, $u_i(x)$ be the feature vector learned from unlabeled data (discussed in Section 3.1), $\boldsymbol{W}$ and $\boldsymbol{V}$ be the corresponding weight matrices, $\boldsymbol{b}$ be the bias vector, and σ be the component-wise non-linear activation function; then a computational unit of the convolutional layer associated with the i-th phrase can be formulated as follows:

$$\sigma(\boldsymbol{W} \cdot r_i(x) + \boldsymbol{V} \cdot u_i(x) + \boldsymbol{b}) \quad (2)$$

$\boldsymbol{W}$, $\boldsymbol{V}$, and $\boldsymbol{b}$ in (2) are the parameters of the model that are learned through training and shared across all neurons of the same layer. The model uses the *rectifier*, $\sigma(x) = max(0, x)$, as the non-linear activation function (ReLU), max-pooling to compute higher-layer abstractions, and stochastic gradient descent for optimization where the objective is to minimize the square loss with respect to the labeled training set. Finally, the output layer of the network uses a linear classifier that exploits the learned features to identify if a tweet describes an ADE or not. For details about the general framework, see [11].

3.4 Experimental Setup

3.4.1 Model Parameters

For training and testing the semi-supervised CNN, we use ConText v2 [11].[16] For all of our semi-supervised CNN models, we use the default parameters of ConText. In particular, for the unsupervised feature embedding learning, we use 1000 neurons, vector dimension of 200, region size of 3, bag of 1-2-3-grams region vector, region vocabulary 30,000, target vocabulary 10,000. For the supervised learning, we set number of iterations (or epochs) to 100, mini-batch size to 1,000, vocabulary size to 10,000.

3.4.2 Training and Test Data

For training the supervised CNN part of the semi-supervised classification and to evaluate the performance of our trained

[13] We suspect, these are annotation errors in the source corpus.

[14] `https://radimrehurek.com/gensim/models/phrases.html`

[15] We additionally experimented with adding tweets from *T-Random* in the combined data sets, but they did not improve classification performance, so we omit these models here.

[16] `http://riejohnson.com/cnn_download.html`

Table 2: Statistics of Evaluation Data

	# of Tweets (Training)	# of Tweets (Test)
Positive instances	628	166
Negative instances	5,052	1,254
Total	5,680	1,420

models, we use the Twitter data set used in an ADR classification work [37]. The data set was created by first collecting tweets with generic and brand names of drugs (in a similar manner our *Model T-Drug* was created), and then a randomly selected sample of the data was annotated by two domain experts under the guidance of a pharmacology expert. The annotators had an inter-annotator agreement of 0.69 (Cohen's Kappa). The authors did not clarify if the annotations distinguished between ADR and ADE, but since ADE cases are a super-set of ADR cases,[2] we use this data set for evaluating our classification models. The same data set was also used in the PSB (Pacific Symposium on Biocomputing) 2016 Social Media Shared Task for ADR classification (Task 1).[1] The tweets in the data set have binary labels for ADR.

The original data set contained a total of 10,822 tweets. As Twitter's terms-of-service do not allow sharing of actual tweet text, the data set is only available via tweet IDs. At the time when we re-acquired the data using the IDs, only 7,100 (65.6%) tweets were still publicly available. We randomly divided the available tweets into training, validation and test data (60%-20%-20% split) for experiments and evaluation. We ensured that none of our unlabeled Twitter data that we used to train the semi-supervised models had any overlap with the evaluation data set (by checking for duplicate text content).

For semi-supervised classification, ConText's [11] default loss function optimizes Mean Squared Error with respect to all classification categories. Because we do a binary classification on a heavily imbalanced class distribution (11% positive and 89% negative), we needed to find an optimal threshold for the prediction probabilities when classifying tweets as ADE. For each of our models, we individually tune the threshold on the validation data so that F1-score for the ADE class is maximized. After the threshold parameter is tuned for each model, we combine the training and validation data to create a final training data set. Table 2 presents statistics of the final training (80%) and test (20%) data sets.

4. EVALUATION

To evaluate the performance of our semi-supervised CNN models, the trained models are applied to each tweet in the 20% held-out test data to predict a binary label (i.e., ADE or non-ADE). We use precision, recall and F1-score as the evaluation metrics and report the results for the ADE class. For comparison, we also present classification performance of a baseline method and a number of supervised classification models. The supervised models are trained on the same final training set (described in section 3.4.2) used in our semi-supervised models.

Drug-Health-Condition Baseline is a simple heuristic based prediction model that classifies a tweet as ADE if the tweet mentions both a drug name and a health condition. For the drug names, we use the dictionary we created from the Walgreen drugs list (described in Section 3.2.3) containing 380 commonly mentioned drug names in Twitter. For the health conditions, we use the dictionary we created (described in Section 3.2.4) containing 7,193 health conditions.

fastText is a fast algorithm recently released by researchers at Facebook [14] that efficiently learns text representations for classification. It achieves comparable to state-of-the-art performance on tasks such as sentiment analysis and tag prediction. The classifier does not use any domain or problem specific features, rather it uses bag-of-n-grams and transforms them into low dimensional vector space so that the features can be shared across classification categories despite their lexical differences. We train fastText on the final 80% training data (with the same vector dimension and epoch we use in our semi-supervised CNN models) and apply the trained model on the final 20% held-out test data for ADE classification.

Supervised CNN is a supervised convolutional neural network classifier trained only on labeled tweets. Since both the supervised and semi-supervised CNN models use the same labeled training instances, we compare their performances to determine if ADE classification results could be improved with additional unlabeled data. For training the supervised CNN, we use ConText with its default parameter settings.

ADR Classifier is a state-of-the-art binary classifier [37] that is designed to classify short texts into ADR or non-ADR categories. It is a supervised classifier that uses a wide range of features derived from n-grams, UMLS semantic types that represent medical concepts, phrases denoting change in sentiment, WordNet synsets, ADR lexicon, sentiment lexicon, topic model-based features, text length, presence of comparative/superlative, modals, etc. The authors reported achieving 53.8% F1-score when the model was trained and tested on the original Twitter data set (10,822 tweets). The authors further improved the classification performance up to 59.7% F1-score on the original Twitter data set by using 10,617 labeled posts from an online health community forum and 23,516 labeled medical case reports as additional training instances. Since it was not possible to re-create the exact training/test data due to unavailability of some tweets, we trained their classification model with our final training data set, and applied it to the tweets in our test data to predict ADE tweets.

4.1 Results

Table 3 presents our ADE classification results and comparisons. The simple *Drug-Health-Condition* baseline results have the highest recall (63.86%) among all classification results, but also the lowest precision (25.67%). This shows that a drug and a health condition co-occurring in a tweet does not necessarily indicate an ADE, since these tweets will also include *indication* events (the drug is used to treat the condition) among other possibilities. The low precision can also be attributed to the noise in the health conditions dictionary.

The next section in Table 3 shows the results for the supervised models. The fastText model, which is a general purpose text classification model, improves performance substantially (+12% F1-score) over the simple *Drug-Health-Condition* baseline, but at the cost of much lower recall. The supervised CNN improves performance further with a substantial recall gain, raising F1-score to 52.53%.

Table 3: ADE Tweet Classification Performance (P = Precision, R = Recall, F1 = F1-score, M = ∼1,000,000 tweets or sentences, K = ∼1,000 tweets or sentences), differences in performance from the last row, majority vote, are statistically significant over the non-shaded regions (p < 0.05)

Model	Unlabeled Data Size	P (%)	R (%)	F1 (%)
Baseline				
Drug-Health-Condition Baseline	NA	25.67	**63.86**	36.61
Supervised Models				
fastText [14]	NA	56.35	42.77	48.63
Supervised CNN	NA	55.33	50.00	52.53
ADRClassifier [37]	NA	55.63	53.62	54.60
Model Built with General Unlabeled Data				
Model 1 T-Random	43.5M	64.29	54.22	58.82
Selective Use of Unlabeled Data (individual data sets)				
Model 2 Sent-Health	484K	58.79	58.43	58.61
Model 3 T-Drug	157K	63.24	51.81	56.95
Model 4 T-Health-Condition	3.5M	64.67	58.43	61.39
Model 5 T-Health-Condition*	7M	65.73	56.63	60.84
Selective Use of Unlabeled Data (multiple data sets)				
Model 6 T-Drug-Condition-Sent-Health	4.2M	67.11	60.24	63.49
Model 7 T-Drug-Condition*-Sent-Health	7.7M	67.33	60.84	63.92
Ensemble Prediction				
Majority Vote	NA	**70.21**	59.64	**64.50**

The state-of-the-art ADR classifier [37] specifically designed for this task achieves the best results (54.60% F1-score) among the supervised models. Despite the difference with the original data set due to unavailability of some tweets, this is still comparable to the reported results in the original paper (53.8% F1-score when trained and tested with Twitter data).

Our results for *Model-1 T-Random*, which uses a massive collection of random Twitter data (43.5M), has substantial performance improvements across all three evaluation metrics when compared to the supervised CNN model, and also outperformed the ADR Classifier baseline. This demonstrates the merits of the semi-supervised CNN classification as it leverages unlabeled data for automatic feature learning, and provides a plausible alternative to supervised methods since human-annotated labeled data are costly to generate or rarely available. However, recall at 54.22%, although better than the ADR Classifier, can still be improved.

In the next section of Table 3, we present the classification results when unlabeled data are used selectively so that the semi-supervised models can learn the problem-specific region embeddings (i.e., for high-level problem concepts) effectively. These models (Model 2-5) use much less unlabeled data than *Model-1 T-Random* and their data set sizes range from only 484K to 7M. Their performances also vary as precision ranges from 58.79% to 65.73%, recall ranges from 51.81% to 58.43% and F1-score ranges from 56.95% to 61.39%. All of these models (Model 2-5) achieve better F1-score than the baseline method and supervised models, while the best performing model's F1-score (*Model-4 T-Health-Condition*) improves over *Model-1 T-Random*'s F1-score by +2.57%. The relative differences in the results for these models (Model 2-5) are hard to compare since their unlabeled data sizes differ from each other. We also do not expect these models to perform substantially better because, individually, their unsupervised learning was tailored to a specific type of high-level concept such as drug names or health conditions. Intuitively, to be able to classify ADE effectively, the model should learn region embeddings of all of these high-level concepts simultaneously.

Model 6 & 7 are semi-supervised CNN models that combine all of the unlabeled data used in Model 2-5, such that region embeddings for drug, health conditions and other possible high-level problem concepts can be learned together. Model-7 outperforms all the models that use individual data sets (Model 2-5), and improves F1-score by an additional +2.53%. This model uses about 7.7M instances, which is nearly 5.6 times fewer number of instances than the random tweets data, and yet, was able to increase F1-score by +5.1% (from *Model-1 T-Random*), reaching 63.92% F1-score. Previous best results [37] reported on this data set had 59.7% F1-score, where the classification model used annotated tweets, online health community forum posts and medical case reports —totalling nearly 40K labeled training instances, whereas in the semi-supervised framework, we could achieve improved classification performance on a comparable data set with only 5,680 labeled training instances and selective use of the unlabeled data.

Since each of our models learns the region embeddings from slightly different types of unlabeled data, in the final row of Table 3 we combine the predictions from different models (Model 2-7) using majority vote for an ensemble prediction method. We break ties using the best performing model's prediction (best F1-score on the validation data set). ADE tweet classification with this ensemble prediction achieves the highest precision (70.21%) and F1-score (64.50%) among all our experiments, outperforming the state-of-the-art supervised ADR classifier from previous work by +9.9% F1-score, while recall level remained similar or better than the individual models (Model 2-7). These final results are statistically significant (using paired bootstrap significance test [2]) for most of the precision and F1-score improvements (P < 0.05), with the exception of precision of Model 6 and F1-score of Model 6 & 7. The non-shaded regions in Table 3 indicate statistically significant differences from the *Majority Vote* prediction results.

Table 4: Examples of False Negatives, False Positives and Prediction Disagreement among Models

Example Tweet	Label	Prediction (Model 6)	Prediction (Model 7)	Majority Vote
False Negatives				
@USER And then I had *horrible sleep* once I took the *trazodone*. I just couldn't win, haha.	ADE	non-ADE	non-ADE	non-ADE
@USER: I run on Vyvanse and RedBull. So done with that life. *Vyvance cooked my brain like a stove top*	ADE	non-ADE	non-ADE	non-ADE
False Positives				
Doctor gave me *moxifloxacin to kill my sinus infection*. Bringing out the big guns; sick of being sick!	non-ADE	ADE	ADE	ADE
@USER depending on your pain, for me and *my back all over pain Cymbalta* has been a miracle *no side effects*	non-ADE	ADE	ADE	ADE
Prediction Disagreement among Models				
Medication side-affects have hit me hard today: I keep on *randomly jolting* or *rocking*, it is so bad I am finding it *hard to type*! *#Seroquel*.	ADE	non-ADE	ADE	ADE

4.2 Qualitative Analysis

Table 4 presents some example tweets for which our best models had prediction errors or there were prediction disagreements among the models.

False Negatives. In the first tweet in Table 4, the health condition 'horrible sleep' is another way of describing 'sleep disorder', but 'horrible sleep' was not in our health conditions dictionary. Both Models 6 & 7 could not recognize this tweet as ADE. The second example may refer to one of the mental side effects of the drug ('manic symptoms', 'bipolar illness', 'Seeing things or hearing voices that are not real', 'Believing things that are not true'), but described in a very informal way ('cooked my brain'). These examples demonstrate the challenge of detecting these medical concepts written in colloquial language.

False Positives. In the middle section of Table 4, we present two examples that our models falsely predicted as ADE. These examples describe an *indication* event where a drug is used to treat a condition. Even though the user specifically mentioned that the drug gives 'no side effect', the model may not have learned good region embeddings for such phrases.

Prediction Disagreement among Models. In the last section of Table 4, we show an example for which our models (model 6 and 7) had a disagreement in their predictions. Model 7 (T-Drug-Condition-Sent-Health) predicted it correctly as an ADE, but model 6 (T-Drug-Condition*-Sent-Health) misclassified it as non-ADE. 'keep on randomly jolting or rocking' and 'hard to type' are not in our health condition dictionary. It is possible that because of other health-related words in the expanded dictionary, Model 7 could recognize the tweet as ADE. The ensemble model that takes a majority vote could also predict the correct label.

5. CONCLUSION

We have presented our experiments with a semi-supervised CNN-based framework for classification of adverse drug events in tweets, and evaluated our models on the Twitter data set used in the PSB 2016 Social Media Shared Task. By leveraging different types of unlabeled data to learn phrase embeddings for the semi-supervised classification, our models outperformed a state-of-the-art ADE classification model by +9.9% F1-score. Our best model (ensemble prediction by majority vote) achieved 70.21% precision, 59.64% recall, and 64.50% F1-score, and to the best of our knowledge sets new state-of-the-art results for this data set. ADE classification in tweets can allow for early detection as a means to augment existing ADE surveillance systems, and our results suggest a feasible solution that does not require a large number of labeled instances. In future, we will explore automatic extraction of high-level ADE concept phrases with the help of learned region embeddings, to detect drug and side-effect associations. Removing noisy phrases from both drugs and health conditions dictionaries could be a possible improvement scope in future work. ADE extraction and colloquial language modeling are additional avenues worth exploring for improving ADE classification performance.

6. REFERENCES

[1] E. Benzschawel. Identifying potential adverse drug events in tweets using bootstrapped lexicons. Master's thesis, Brandeis University, 5 2016.

[2] T. Berg-Kirkpatrick, D. Burkett, and D. Klein. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

[3] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270, 2004.

[4] E. G. Brown, L. Wood, and S. Wood. The medical dictionary for regulatory activities (meddra). *Drug Safety*, 20(2):109–117, 2012.

[5] H.-J. Dai, M. Touray, J. Jonnagaddala, and S. Syed-Abdul. Feature engineering for recognizing adverse drug reactions from twitter posts. *Information*, 7(2):27, 2016.

[6] C. N. dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, 2014.

[7] D. Egger, F. Uzdilli, M. Cieliebak, and L. Derczynski. Adverse drug reaction detection using an adapted sentiment classifier. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.

[8] C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, and N. Dasgupta. Digital drug safety surveillance: Monitoring

pharmaceutical products in twitter. *Drug Safety*, 37(5):343–350, 2014.

[9] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, pages 885 – 892, 2012.

[10] S. A. Hasan, Y. Ling, J. Liu, and O. Farri. Exploiting neural embeddings for social media data analysis. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, 2015.

[11] R. Johnson and T. Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Proceedings of the 29th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2015.

[12] R. Johnson and T. Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. *arXiv preprint arXiv:1602.02373*, 2016.

[13] J. Jonnagaddala, T. R. Jue, and H. Dai. Binary classification of twitter posts for adverse drug reactions. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Big Island, HI, USA*, 2016.

[14] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[15] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[16] N. Kang, B. Singh, C. Bui, Z. Afzal, E. M. van Mulligen, and J. A. Kors. Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics*, 15(1):1, 2014.

[17] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73 – 81, 2015.

[18] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[19] J. Lardon, R. Abdellaoui, F. Bellet, H. Asfari, J. Souvignet, N. Texier, M. C. Jaulent, M. N. Beyens, A. Burgun, and C. Bousquet. Adverse drug reaction identification and extraction in social media: A scoping review. *Journal of Medical Internet Research*, 17(7):e171, 2015.

[20] R. Leaman, R. I. Dogan, and Z. Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[21] R. Leaman, R. Khare, and Z. Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37, 2015.

[22] Y. Lecun and Y. Bengio. *Convolutional Networks for Images, Speech and Time Series*. The MIT Press, 1995.

[23] J. Y. Lee and F. Dernoncourt. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[24] K. Lee, A. Agrawal, and A. Choudhary. Real-time disease surveillance using twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1474–1477, New York, NY, USA, 2013. ACM.

[25] K. Lee, A. Agrawal, and A. Choudhary. Mining social media streams to improve public health allergy surveillance. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 815–822, Aug 2015.

[26] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258, Dec 2011.

[27] N. Limsopatham and N. Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[28] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *ACL*, 2011.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.

[30] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[31] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22:671–681, 2015.

[32] B. Ofoghi, S. Siddiqui, and K. Verspoor. Read-biomed-ss: Adverse drug reaction classification of microblogs using emotional and conceptual enrichment. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.

[33] V. Plachouras, J. L. Leidner, and A. G. Garrow. Quantifying self-reported adverse drug events on twitter: Signal and topic analysis. In *Proceedings of the 7th 2016 International Conference on Social Media & Society*, 2016.

[34] M. Rastegar-Mojarad, R. K. Elayavilli, Y. Yu, and H. Liu. Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.

[35] H. Sampathkumar, X. Chen, and B. Luo. Mining adverse drug reactions from online healthcare forums

using hidden markov model. *BMC Medical Informatics and Decision Making*, 14(1):1–18, 2014.

[36] A. Sarker, R. E. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212, 2015.

[37] A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196 – 207, 2015.

[38] D. Tang, B. Qin, and T. Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[39] Unified medical language system Umls metathesaurus. `https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html`. [accessed September-2016].

[40] S. J. Yeleswarapu, A. Rao, T. Joseph, V. Saipradeep, and R. Srinivasan. A pipeline to extract drug-adverse event. *BMC Med. Inf. & Decision Making*, 14:13, 2014.

[41] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, 2015.

[42] Z. Zhang, J.-Y. Nie, and X. Zhang. An ensemble method for binary classification of adverse drug reactions from social media. *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.