Learning Causality Patterns for Detecting Adverse Drug Reactions from Social Media

Danushka Bollegala, Richard Sloane, Simon Maskell, Joanne Hajne, Munir Pirmohamed

Corresponding author:
Danushka Bollegala
Email: danushka.bollegala@liverpool.ac.uk
Phone: +44-1517954283
University of Liverpool, Liverpool, L693BX, United Kingdom.

Abstract

Background: Detecting adverse drug reactions (ADRs) is an important task that has direct implications for the use of that drug. If we can detect previously unknown ADRs as quickly as possible, then this information can be fed back to regulators, pharmaceutical companies and healthcare organisations thereby potentially reducing drug-related morbidity and saving lives of many patients. A promising approach for detecting ADRs is to use social media platforms such as Twitter and Facebook. A high level of correlation between a drug name and an event may be an indication of a potential adverse reaction associated with that drug. Although numerous association measures have been proposed by the signal detection community for identifying ADRs, these measures are limited in that they detect correlations but often ignore causality.

Objective: In this paper, we propose a causality measure that can detect an adverse reaction that is caused by a drug rather than merely being a correlated signal.

Methods: To the best of our knowledge, this is the first causality-sensitive approach for detecting ADRs from social media. Specifically, we represent the relationship between a drug and an event using a set of automatically extracted lexical patterns. We then learn the weights for the extracted lexical patterns that indicate their reliability for expressing an adverse reaction of a given drug.

Results: Our proposed method obtains an ADR detection accuracy of 74% on a large-scale manually annotated dataset of tweets, covering a standard set of drugs and adverse reactions.

Conclusions: By using lexical patterns, we can accurately detect the causality between drugs and adverse reaction related events.

Trial Registration: This study does not require any clinical trials.

Keywords: ADR detection; causality; lexical patterns; causality detection; support vector machines

Introduction

An Adverse Drug Reaction (ADR) is defined as "an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, alternation of the dosage regimen, or withdrawal of the product" [34, 15, 26, 1]. It is estimated that approximately 2 million patients in the United States are affected each year by serious ADRs, resulting in roughly 100,000 fatalities [21]. In fact, ADRs are the fourth leading cause of death in the U.S. following cancer and heart diseases [19]. Treating patients who develop ADRs results in significant health costs to nations throughout the world. For example, in the U.S. it has been estimated that USD 136 billion is spent each year on treatments related to ADRs [6, 32].

In an ideal world, all adverse reactions associated with a drug need to be detected prior to marketing, and the drug label modified accordingly. However, this is not feasible due to several reasons. First, the number of human subjects participating in a clinical trial of a pre-marketed drug is often small, which limits the statistical power to detect ADRs, particularly those which may be uncommon. In fact, rare ADRs are usually not detected during the pre-marketing phases of drug development. Second, since many of the clinical trials are short-lasting, ADRs which are delayed will not be detected. Third, some ADRs show up only when a drug is being taken together with other drugs leading to an adverse drug-drug interaction. Considering that the number of combinations of drugs is potentially large, it is impractical to test for all of the possible combinations during a clinical trial. Fourth, *drug repurposing* [28] – the practice of off-label usage of drugs for treating diseases for which they were not originally intended, could lead to unforeseen ADRs.

Because of these challenges in detecting ADRs during the pre-marketing phase, identification of ADRs in the post-marketing phase remains hugely important. The cornerstone of post-marketing pharmacovigilance remains the spontaneous reporting schemes such as the Yellow Card Scheme [35] in the UK and the MedWatch system [36] in the US. Such schemes allow hospitals, medical practitioners, and patients to report ADRs. Unfortunately, the reporting rates are generally poor. For example, only 10% of serious ADRs and 2–4% of nonserious ADRs are reported [9].

Although patients experience ADRs, they may be reluctant to report their experiences through official reporting systems for various reasons. For example, patients might be unfamiliar with or unaware of the ADR reporting schemes, or might find it difficult to understand the terminology used in the forms, or might not be aware of the importance of reporting ADRs. Even when ADRs have been reported via such spontaneous reporting systems, the time required from the first report to any regulatory action may be long, which is problematical in protecting public health from iatrogenic conditions.

An alternative approach for detecting ADRs in a timely manner on a larger scale is to use social media. Social media platforms such as Twitter [37], Facebook [38], Instagram [40], Pinterest [39] etc. have been used extensively for market analysis of various products. Social media provides a convenient and direct access to consumers' opinions about the products and services they use. In comparison to a clinical study, which inevitably is limited to a small number of participants, in social media we can access comments from a massive number of diverse groups of people. Because of its potential value, the pharmacovigilance community has already started to exploit social media as a potential reporting tool for obtaining information about ADRs [30]. For example, the WEB-RADR [41] project funded by the Innovative Medicines Initiative (IMI) was funded to evaluate the usefulness of social media as a reporting tool for ADRs.

However, compared to spontaneous reporting systems where patients or healthcare practitioners explicitly report ADRs, detecting ADRs from social media poses several challenges. Because social media is not perceived by most patients as an official reporting tool for ADRs, a drug and its associated ADRs might not be completely expressed in a single social media post. This issue is further aggravated by the limitations imposed on the length of a post in social media platforms. For example, in Twitter, a single post (aka. a *tweet*) is limited to a maximum of 140 characters. Even in social media platforms where such limitations do not exist such as Facebook, the users might not always provide comprehensive reports containing all the information that would normally be completed on a Yellow Card. Furthermore, social media users often interact with social media platforms through specialised apps on mobile devices such as smart phones, which do not possess physical key boards that facilitate the entering of longer texts.

In addition to the brevity and incompleteness of social media posts as a medium for reporting ADRs, the reliability of the information expressed through social media is also a concern. It is often difficult to authenticate the information disseminated through social media. For example, in Twitter, the same user can create multiple accounts under different names including aliases. False information might be expressed intentionally or unintentionally in social media, which makes it difficult to verify the information extracted from social media. Unlike in the Yellow Card system, where it is possible to contact a reporter to obtain further information, in social media it is difficult to obtain additional information from users due to anonymity and privacy settings. All of these challenges introduce various levels of noise to ADR signal that can be captured from social media. Consequently, methods that detect ADRs from social media need to overcome these challenges.

An approach for detecting significant signals indicating adverse reactions to drugs in social media is to measure the correlation between a drug and an event. If many social media posts and/or users mention a drug and an event, then the likelihood that the drug causing an adverse reaction increases. Indeed, numerous measures have been proposed in previous work to measure the degree of association between a drug and an adverse reaction [31, 24, 5, 4, 11, 2, 17, 12]. Although co-occurrence measures do not completely solve all of the above-mentioned challenges of using social media, they provide a practical and a highly scalable mechanism for detecting ADRs from social media.

A fundamental drawback of co-occurrence-based approaches for detecting ADRs is that they ignore the context in which a drug and an ADR co-occur in social media. Co-occurrence *does not* always indicate causality. Although a drug and an event which could suggest an ADR might be mentioned frequently in social media, the co-occurrence may be because the drug is used as a remedy for that symptom. Moreover, the drug may have been taken by one person but the social media post mentions the ADR in a different person. However, the context in which a drug and an ADR co-occur can provide useful clues that can be used to separate causality from co-occurrence. The co-occurring context between a drug and an ADR provides useful clues that we can use to separate causality from co-occurrence.

To illustrate the usefulness of contextual information for ADR detection consider the three tweets shown in Figure 1.

T_1: **Benlysta** week two. Injections in. About to *fall asleep* real quick.

T_2: @Mariah2you I feel really good. I still have random things happen (*rashes*, *tear duc blockage*) but im down to 5mg of **prednison** and **Benlysta**

T_3: Pop some **ibuprofen** and pass out #*sleep* #*exhausted*

Figure 1: Three tweets mentioning a drug (shown in blue boldface fonts) and an symptoms (shown in red italic font).

T_1 is suggestive of an association with a drug and a potential adverse reaction. T_2 may reflect that the patient's disease improving or that an ADR occurred but is waning following dose reduction. T_3 is unlikely to be an ADR; Ibuprofen is being taken by this patient to potentially relive the pain and have some sleep. These examples show that there are useful hints we can extract from the tweets such as *about to (feel an ADR)*, *I still have (ADRs)* that we can use to evaluate the causality relationship between a mentioned drug and an adverse reaction.

Why is solving this problem critical for systems that attempt to extract ADRs from social media? The standard practice in the pharmacovigilance community for detecting ADRs from patient reports is to apply disproportionality measures that consider *only* co-occurrence (and occurrence) counts. Unfortunately, disproportionality measures by design are agnostic to the linguistic context in social media, and are therefore unable to utilise the clues that appear in social media to determine whether an ADR is truly caused by the drug. However, given a tweet containing a drug and a potential adverse reaction, if we can first develop a classifier that predicts whether this tweet is describing a causality relationship, we then can use disproportionality measures on the tweets that are identified as positive by the classifier for further analysis. This pre-processing step is likely to improve the accuracy of the ADR detection process. Moreover, given the noise and the low-level of reliability in social media as opposed to patient reports in spontaneous reporting schemes, it is vital that we perform some form of pre-processing to guarantee the reliability of the identified ADRs.

In this paper we therefore consider the following problem: given a tweet T containing a drug D and an ADR A, whether T describes an instance where A is caused by D, as opposed to A and D co-occurring for a different reason (or randomly without any particular relation between A and D). Our experimental results show that the proposed method statistically significantly outperforms several baseline methods, demonstrating its ability to detect causality between drugs and ADRs in social media.

Related Work

The number of co-occurrences between a drug and an ADR can be used as a signal for detecting ADRs associated with drugs. Various measures have been proposed in the literature that evaluate the statistical significance of disproportionally large co-occurrences between a drug and an ADR. These includes (M)GPS ((Multi-item) Gamma Poisson Shrinker) [13, 12, 20, 2], RGPS (Regression-Adjusted Gamma Poisson Shrinker) [11], BCPNN (Bayesian Confidence Propagation Neural Network) [4, 5, 24], PRR (Proportional Reporting Rate) [20, 31], and ROR (Reporting Odds Ratio) [20, 31]. Each of these algorithms uses a different measure of disproportionality between the signal and its background. Information component (IC) is applied in BCPNN, while empirical Bayes geometric mean is implemented in all variants of the GPS algorithm. Each of the measures gives a specific score, which is based on the number of reports including the drug or the event of interest. These count-based methods are collectively referred to as *disproportionality measures*.

In contrast to these disproportionality measures which use only co-occurrence statistics for determining whether there is a positive association between a drug and an event, in this paper, we propose a method that uses the contextual information extracted from social media posts to learn a classifier that determines whether there is a causality relation between a drug and an ADR. Detecting causality between events from natural language texts has been studied in the context of discourse analysis [10, 27] and textual entailment [3, 18]. In discourse analysis, a discourse structure for a given text is created showing the various discourse relationships such as causality, negation, evidence etc. For example, in Rhetorical Structure Theory (RST) [22], a text is represented by a discourse tree where the nodes correspond to sentences or clauses referred to as Elementary Discourse Units (EDUs), and the edges that link those textual nodes represent various discourse relations that exist between two EDUs. Supervised methods that require manually annotated discourse trees [14] as well as unsupervised methods that use discourse cues [23] and topic models [25] have been proposed for detecting discourse relations.

The problem of determining whether a particular semantic relation exists between two given entities in a text is a well-studied problem in the NLP community. The context in which two entities co-occur provide useful clues for determining the semantic relation that exists between those entities. Various types of features have been extracted from co-occurring contexts for this purpose. For example, Cullotta and Sorensen [52] proposed tree kernels that use dependency trees. Dependency paths and the dependency relations over those paths are used as features in the kernel. Agichtein and Gravano [53] used a large set of automatically extracted surface-level lexical patterns for extracting entities and relations from large text collections.

To address the limitations of co-occurrence-based approaches, several prior work have used contextual information [46]. Nikfarjam et al. [45] annotated tweets for ADRs, beneficial effects and indications, and used those tweets to train a Conditional Random Field (CRF). They use contextual clues from tweets and word embeddings as features. Their problem setting is different from ours in the sense that we do not attempt to detect/extract ADRs or drug names from tweets but are only interested in determining whether the mentioned ADR is indeed relevant to the mentioned drug. A tweet can mention an ADR and a drug but the ADR might not necessarily be related to the ADR. Huynh et al. [47] proposed multiple deep learning models by concatenating convolutional and recurrent neural network architectures to build ADR classifiers. Specifically, given a sentence, they would like to create a binary classifier that predicts whether the sentence contains an ADR or otherwise. Their experimental results show that convolutional neural networks to be the best for ADR detection. This observation is in agreement with broader text classification tasks in NLP where convolutional neural networks have reported the state-of-the-art performance [48]. However, one issue when using CNNs for ADR detection is the lack of labelled training instances, such as annotated tweets. This problem is further aggravated if we must learn embeddings of novel drugs or rare ADRs as part of the classifier training.

To overcome this problem, Lee et al. [49] proposed a semi-supervised convolutional neural network that can be pretrained using unlabeled data for learning phrase embeddings. Bidirectional Long Short-Term Memory (bi-LSTM) units were used in [50] to tag ADRs and indicators in tweets. A small collection of 841 tweets were manually annotated by two annotators for this purpose. Pretrained word embeddings using skip-gram on 400 million tweets are used to initialise the bi-LSTM's word representations. This setting is different to what we study in this paper because we do not aim to tag ADRs and indicators in a tweet but to determine whether a tweet that mentions an ADR and a drug indicator describes an ADR event related to the drug mentioned in the tweet.

Methods

In this section, we present our proposed method for detecting the causality between a drug and an event. First, in Section **Error! Reference source not found.**, we formally define the problem of causality detection between a drug and an event from social media posts. Second, in Section **Error! Reference source not found.**, we explain techniques for aggregating social media posts related to drugs and events. Third, in Section **Error! Reference source not found.**, we explain the method we use for extracting various lexical patterns that describe the relationship between a drug and an event in social media posts. Finally, in Section **Error! Reference source not found.**, we present a machine learning approach that uses a manually annotated dataset containing social media posts as to whether they are describing a relationship between a drug and an adverse reaction for learning the reliability of the lexical patterns we extract in Section **Error! Reference source not found.**. We do not assume any specific properties or meta-data available in a particular type of social media platform such as *retweets*, *favourites* in Twitter, or *likes* or *comments* in Facebook. Although such platform-specific meta-data can provide useful features for a machine learning algorithm, such meta-data are not universally available across all social media platforms or cannot be retrieved due to privacy settings. The fact that the proposed method does not rely on such meta-data is attractive because it makes our proposed method applicable to a wide-range of social media posts, and does not limit it to a particular platform.

Problem Definition

Let us consider a social media post T, which explicitly mentions a drug D and an adverse reaction R. We model the problem of detecting causality between D and R in T as a binary classification problem where we would like to learn a binary classifier $h(T,D,R;\boldsymbol{w})$ parametrised by a d-dimensional real-valued weight vector $\boldsymbol{w}\in\mathbb{R}^d$ as follows:

$$h(T,D,R;\boldsymbol{w}) = \begin{cases} 1, & \text{if } T \text{ mentions that } D \text{ causes } R \\ 0, & \text{otherwise0} \end{cases} \qquad (1)$$

Here, we assume that the social media post T is already given to us and the drug and adverse reaction have already been detected in T. Detecting drug names can be done by matching against pre-compiled drug name lists (gazetteers) or using Named Entity Recognition (NER) tools [29]. A particular challenge when matching drug names in social media is that the drug names mentioned in social media might not necessarily match against the drug names listed in pharmacology databases [30]. The same drug is often sold under different labels by different manufacturers, and the label names continuously change, which makes it difficult to track a particular drug over time in social media. Similar challenges are encountered when matching ADRs in texts. Although the MedDRA [42] hierarchy assigns unique codes to preferred terms (PTs) that describe various ADRs such as “oropharyngeal swelling” or “systemic inflammatory response syndrome”, such terms are used rarely by the majority of the social media users who might not necessarily be familiar with the MedDRA code names [55]. Although we acknowledge the challenges in detecting mentions of drug names and adverse reactions, we consider it to be beyond the scope of the current paper, which focuses on a signal detection problem.

Social Media Aggregation

Although the problem definition described in Section **Error! Reference source not found.** assumes that we are already provided with a set of social media posts, obtaining a large collection of social media posts relevant to drugs and events can be challenging for several reasons.

The vast majority of social media posts are not relevant to drugs or ADRs. One effective method for filtering out such irrelevant social media posts is to use the keyword-based filtering functionalities provided by the major social media APIs. As a specific example of such an API, we discuss the use of Twitter streaming API [43]. the Twitter streaming API allows registration of a set of keywords and if there are any tweets that contain at least one of those keywords, then the corresponding tweet will be filtered and sent to the querying user. In our case, we used drug names and PTs (and their lexical variants) as keywords to filter the relevant tweets. Moreover, the streaming API also enabled us to limit the tweets to a particular geographical area or a language, which is useful if we want to monitor drugs that are specifically used in a particular country or a region.

Twitter's streaming API allowed us to aggregate tweets from two main types of data streams: *public streams* and *user streams*. Public streams are publicly available tweets by a specific group of users or on a topic. Hash tags in twitter are useful for streaming such public tweets on a particular topic. For example, by including the hash tag *#epilepsy*, we can retrieve tweets that are relevant to epilepsy. On the other hand, user streams allows us to obtain tweets from a single twitter user, containing roughly all of the data corresponding with that user's view (timeline) on Twitter. Despite the used aggressive filtering, streaming API returned a large number of tweets. Therefore, we stored the filtered tweets in a MongoDB [44] database in JSON format for efficient retrieval.

Lexical Pattern Extraction

To represent the relationship between a drug and an ADR in a tweet, we extracted lexical patterns from the tweet. Let us illustrate the lexical pattern extraction process using the example tweet shown in Figure 2. We first identified the drug and event in the tweet and split the tweet into three parts. The part from the beginning of the tweet to the first mentioned entity (either the drug or event) is named as the *prefix*, the part from the first mentioned entity to the second mentioned entity is named as the *midfix*, and the part from the second mentioned entity to the end of the tweet is named as the *postfix*. Prior work on information extraction has shown that, in English, the midfix provides useful clues related to the relationship between two entities that co-occur in some context [8, 7]. Indeed, from the example shown in Figure 2 we see that words such as *feeling* that appear in the midfix indicate that this twitter user is experiencing a side effect from the drug. However, it has also been shown that prefix and postfix terms also provide useful information when determining the relationship between two entities. For example, we see that the word *took* that appears in the prefix in the tweet (Figure 2) indicating that this twitter user has indeed taken this drug and not simply reporting an adverse reaction experienced by a different person. Such information is useful to estimate the reliability of the relationships mentioned in social media, which can often be noisy and unreliable. Therefore, in this work, we use all prefix, midfix, and postfix sections in tweets for extracting lexical patterns. We experimentally evaluate the significance of prefix, midfix, and postfix for ADR detection later in Section **Error! Reference source not found.**.

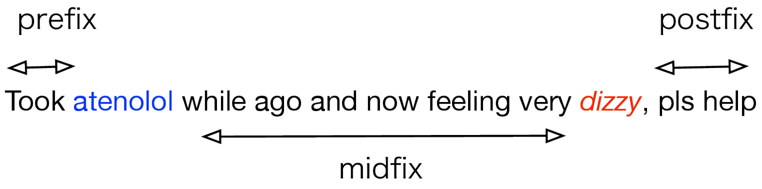
prefix
postfix
Took atenolol while ago and now feeling very dizzy, pls help
midfix

Figure 2: Extracting lexical patterns from a tweet that describe the an adverse reaction (dizziness) caused by a drug (Atenolol). The tweet is split into three parts, *prefix*, *midfix*, and *postfix*, and various lexical patterns are extracted from each part. See text for the details of the pattern extraction method. Best viewed in colour.

We extracted skip-grams from prefix, midfix and postfix separately as lexical patterns for representing the relationship between a drug and an event. A skip-gram is an extension of *n*-gram. Unlike, *n*-grams that require us to consider all consecutive *n* words in a sequence, skip-grams allow us to generalise the *n*-gram patterns by skipping one or more words in a sequence. For example, a trigram (*n*=3) lexical patterns extracted from the midfix shown in Figure 2 would be *while ago and*, *ago and now*, *and now feeling*, *now feeling very*.

On the other hand, skip-gram patterns also let us match any word (indicated by the wildcard "*") in an *n*-gram pattern. For example, the skip-gram pattern ** ago*, which is a generalisation of the bigram pattern *while ago* will match various other time indicators such as *hours ago*, *days ago*, and *months ago*. Unlike, *n*-gram patterns that might not match exactly in numerous other tweets, skip-gram patterns flexibly match different tweets, thereby leading to a dense feature space. More importantly, skip-gram patterns subsume *n*-gram patterns. Therefore, all tweets that can be represented using *n*-gram patterns can be matched by the corresponding skip-gram patterns.

Considering the fragmented, ungrammatical, misspelled texts frequently encountered in social media, skip-gram lexical patterns provide a robust and flexible feature representation. Moreover, extracting skip-grams is computationally efficient compared to, for example, part-of-speech tagging or dependency parsing social media, considering the volume of the texts we must process. Note that the drug name or the event are *not* part of the skip-gram lexical patterns. In other words, we replace the drug name and event respectively by place holder variables D and R. This is important because we would like to generate patterns that not only match the existing drugs and adverse reactions but can generalise to future drugs and their (currently unknown) adverse reactions. In our experiments, we use skip-gram lexical patterns for n=1,2 and 3 and allowed a maximum of one wildcard in a pattern.

Learning Pattern Weights

We built a binary classifier that could predict whether an event R mentioned in a tweet T alongside with a drug D was actually related to D. As explained later in Section **Error! Reference source not found.**, we used a manually annotated collection of tweets where each tweet contained a drug and an event, and a human annotator annotates whether the mentioned ADR is relevant to the drug (positively labeled instance) or otherwise (negatively labelled instance). We represent a tuple (T,D,R) using a feature vector $\boldsymbol{\varphi}(T,D,R) \in \mathbb{R}^{\mathrm{d}}$, where each dimension corresponds to a particular skip-gram lexical pattern we extracted following the procedure described in Section **Error! Reference source not found.**. The value of the i-th dimension in the feature vector is set to 1 if the skip-gram lexical pattern l_i appears in T, or zero otherwise. In other words, each tuple (T,D,R) is represented by a boolean-valued feature vector over the set of skip-gram lexical patterns we extracted from all of the training instances. Using the above notation, let us denote this training dataset by D_{train} = $\{(\boldsymbol{\varphi}(T_n,D_n,R_n)\ ,\mathrm{y}_{\mathrm{n}})\}^{\mathrm{N}}_{\mathrm{n=1}}$. Here, (T_n,D_n,R_n) indicates the n-th training instance out of N total instances in the dataset, and $y_n \in \{-1,+1\}$ indicates the manually annotated label to the n-th instance.

Unfortunately, not all skip-gram lexical patterns are equally important when determining whether there exists a relationship between a drug and an event. For example, in Figure 2, the pattern *while ago* can appear in various contexts, not necessarily in the context where an adverse reaction is described. Therefore, we assigned some form of a *confidence weight* to each skip-gram pattern before we used those patterns to make a decision about the relationship between a drug and an event. For this purpose, we assigned a weight $w_i \in$ to each skip-gram lexical pattern l_i. We then predicted the relationship between D and R in T using the linear binary classifier given by (2).

$$h(T,D,R;\boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(T,D,R) \qquad (2)$$

Here, $\mathbf{w} \in \mathbb{R}^{\mathrm{d}}$ is a d-dimensional real-valued weight vector where the i-th dimension represents the confidence weight w_i we have on the skip-gram lexical pattern l_i as a reliable indicator of a positive relationship between D and R in T. The sign function, sgn, is defined in (3), which returns the sign of the inner-product between the weight vector and the feature vector.

$$\mathrm{sgn}(\theta) = \begin{cases} 1, & \theta > 0 \\ -1, & \theta \leq 0 \end{cases} \qquad (3)$$

Given the training dataset D_{train}, our goal was to learn $\boldsymbol{w}$ such that it can be used in (2) to predict whether the R mentioned in a T with D was indeed related to D. For this purpose, we used linear kernel Support Vector Machines (SVMs) [33] with slack variables $\xi_n \geq 0$. Slack variables act in two ways during training. First, slack variables can be used to absorb the labelling noise in training instances. Given the scale of the annotation task, it is unavoidable that some of the instances will be incorrectly labelled by the human annotators, introducing some labelling noise to the training dataset. Second, slack variables can shift some of the training instances closer to the decision hyperplane, thereby artificially making the dataset to be linearly separable.

Although non-linear kernels such as polynomial, radial basis function (RBF), or sigmoid can be used with SVMs, we limited our analysis to linear kernels for the following reason. Under the linear kernel, the weight associated with a particular feature can be seen as the influence imparted by that feature on the classification decision. This property is useful because we can identify the most discriminative lexical patterns that indicate a positive association between a drug and an event. We can use such lexical patterns, for example, to create extraction rules in the form of regular expressions to extract adverse reactions of drugs from social media. Because we are using a linear classifier in this work, it is important to handle the instances that violate the decision hyperplane using slack variables.

The joint learning of slack variables and weights can be formulated as the constrained convex optimisation problem given by (4).

$$\begin{aligned} \text{minimise} \quad & \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\textstyle\sum_{n=1}^{N}\xi_n \\ & y_n \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(T_n, D_n, R_n) \geq 1 \\ & \xi_n \geq 0 \end{aligned} \qquad (4)$$

Here, C>0, cost factor, is a hyperparameter that determines how much penalty we assigned to margin violations. The optimisation problem given in (4) can be converted into a quadratic programming problem by introducing Lagrange multipliers. Efficient implementations that scale well to large datasets with millions of instances and features have been proposed [16].

Once we have obtained the weights w_i for the skip-gram lexical patterns, we can use (2) to predict the relationship between D and R in T.

Results

We trained and evaluated the proposed method using a manually annotated dataset. The details of the dataset are presented in Section **Error! Reference source not found.**. Next, to evaluate the proposed method we compared it against several baseline methods. The baseline methods and their performances are described in Section **Error! Reference source not found.**.

Dataset and Evaluation Measure

To create a training and testing dataset for our task, we manually annotated a set of social media posts collected from the Twitter and Facebook between the period of August-October 2015. Using the social media aggregation techniques described in Section **Error! Reference source not found.**, we filtered social media posts that contained a single mention of a drug and an event. The number of tweets that contain both a PT and a drug name was 94,890.

We then asked a group of annotators, who are familiar with ADRs of drugs, to annotate whether the event mentioned in the social media post is caused by the drug mentioned in the same post (a positively labelled instance) or otherwise (a negatively labelled instance).

The final annotated dataset contained 44,809 positively labelled instances and 50,081 negatively labelled instances. We perform 5-fold cross-validation on this dataset, selecting 80% of the positive and negative instances in each fold as training data, and the remainder as the testing data. In addition to the above mentioned social media posts, we set aside 1000 positively and 1000 negatively labelled social media posts as developmental data, for tuning the hyperparameter C. In total, we extracted 168,663 skip-gram patterns from this dataset. We used classification accuracy defined by (5) as the evaluation measure.

$$\text{Classification Accuracy} = \frac{\text{Total no. of correctly predicted instances}}{\text{Total no. of instances in the dataset}} \qquad (5)$$

Discussion

We compared the proposed method against several baseline methods using the classification accuracy on the testing data as shown in Table 1. Next, we describe the different methods compared in Table 1.

Majority Baseline: Note that our training and test datasets were unbalanced in the sense that we have more negatively labelled instances than positively labelled instances. This situation is natural given that most social media posts might not necessarily describe an adverse reaction of a drug even though it mentioned both the drug and an event. The training and test datasets we used in our evaluations closely simulate this situation. However, if a dataset is unbalanced, then by simply predicting the majority class (in our case this is the negative label) can still result in classification accuracies greater than 50%. The majority baseline shows the level of performance that was obtained by such a majority classifier.

Bag-of-words Classifier: Our proposed method used skip-gram patterns for representing social media posts. An alternative approach would be to ignore the word order in the text, and represent a text using the set of words contained in it. Specifically, we would represent each text by a binary-valued feature vector where the feature values for the unigrams that appear in the text are set to 1, and 0 otherwise. We then trained a binary SVM classifier with a linear kernel. By comparing against the bag-of-words classifier, we can empirically evaluated the usefulness of the proposed skip-gram lexical patterns.

Prefix only: This is a scaled-down version of the proposed method that used skip-gram patterns extracted only from the prefix. By evaluating against the prefix only baseline, we evaluated the importance of the information contained in the prefix.There are 50021 prefix skip-gram patterns in total.

Midfix only: This is a scaled-down version of the proposed method that uses skip-gram patterns extracted only from the midfix. By evaluating against the midfix only baseline, we evaluated the importance of the information contained in the midfix. There are 53057 midfix skip-gram patterns in total.

Postfix only: This is a scaled-down version of the proposed method that uses skip-gram patterns extracted only from the postfix. By evaluating against the postfix only baseline, we evaluated the importance of the information contained in the postfix. There are 65585 postfix skip-gram patterns in total.

Prefix+Midfix: In this baseline method we used both prefix and midfix for extracting skip-gram patterns. This baseline demonstrates the effectiveness of combining contextual information from both the prefix and the midfix.

Prefix+Postfix: In this baseline method we used both prefix and postfix for extracting skip-gram patterns. This baseline demonstrates the effectiveness of combining contextual information from both the prefix and the postfix.

Midfix+Postfix: In this baseline method we use both midfix and postfix for extracting skip-gram patterns. This baseline demonstrates the effectiveness of combining contextual information from the midfix and the postfix.

CNN: We use the state-of-the-art short text classification method proposed by Kim [51] to train an ADR classifier. Each word in a tweet are represented using 128 dimensional word embeddings, where each dimension is randomly sampled from a uniform distribution in range [-1,1]. The word embeddings are concatenated to represent a tweet. Next, a one-dimensional convolutional neural network (CNN) with a stride size of 3 tokens and a max pooling layer is applied to create a fixed 20 dimensional tweet representation. We use AdaGrad [56] for optimisation with initial learning rate set to 0.01 and the maximum number of iterations is set to 1000. Finally, logistic sigmoid unit is used to produce a binary classifier.

Proposed Method: This is the method proposed in this paper. We use prefix, midfix, and postfix for extracting skip-gram patterns.

Using the development data we found the cost parameter C for each setting. For the bag-of-words classifier the optimal C value was found to be 0.01, whereas for all the variants of the proposed method it was 1.0.

The classification accuracies obtained for the 5-fold cross-validation task for the above-mentioned methods are shown in Table **1**. From Table **1**, we see that the majority baseline achieves an accuracy of 63.19%. Our task here is binary classification and to compute confidence intervals for accuracies we must compute Binomial confidence intervals. There are several ways to compute this and one approach is the use of Clopper-Pearson confidence intervals [54]. By using confidence intervals, we can easily compare the statistical significance between methods, without having to conduct numerous pairwise comparisons between different methods. We compared all other methods against the accuracy reported by the majority baseline using Clopper-Pearson confidence intervals (p<0.001) to test for statistical significance, which is [61.70,65.65]. Statistically significant accuracies over the majority baseline are indicated by an asterisk in Table **1**.

Table 1: Classification accuracy of different baselines and the proposed method. (* indicates statistically significant values)

Method	**Classification Accuracy**
Majority Baseline	63.19
Bag-of-words Classifier	69.31^*
CNN	69.26^*
Prefix only	66.41^*
Midfix only	72.78^*
Postfix only	68.08^*
Prefix+Midfix	74.72^*
Prefix+Postfix	71.07^*
Midfix+Postfix	77.10^*
Proposed method	$\mathbf{77.70}^*$

From Table 1 we see that the best performance is obtained by the proposed method using the skip-gram patterns extracted from all prefix, midfix, and suffix contexts. A skip-gram pattern is an extension of n-gram patterns. Unlike n-gram patterns that must contain consecutive tokens, skip-gram patterns can skip one or more tokens when representing a subsequence. Among the different context types, we see that midfix performs best, whereas prefix and postfix performs relatively equally. This result is in agreement with prior work on information extraction for English, where midfix has been found to be useful. However, to the best of our knowledge, such an analysis has not yet been conducted for ADR extraction. Interestingly, we see that by adding the midfix to prefix and postfix we always perform better than if we had used only prefix or postfix. The proposed method uses all three contexts and obtains the best performance among the methods compared in Table 1. In particular, the performance reported by the proposed method is statistically significant over both the majority baseline and the bag-of-words classifier. We see that the CNN-based ADR classifier is performing at the same level as the BOW classifier. Compared to the typical sentence classification datasets used to train such deep learning methods, our twitter dataset is significantly smaller and this lack of data might have resulted in CNN-based ADR classifier to perform poorly in our experiments.

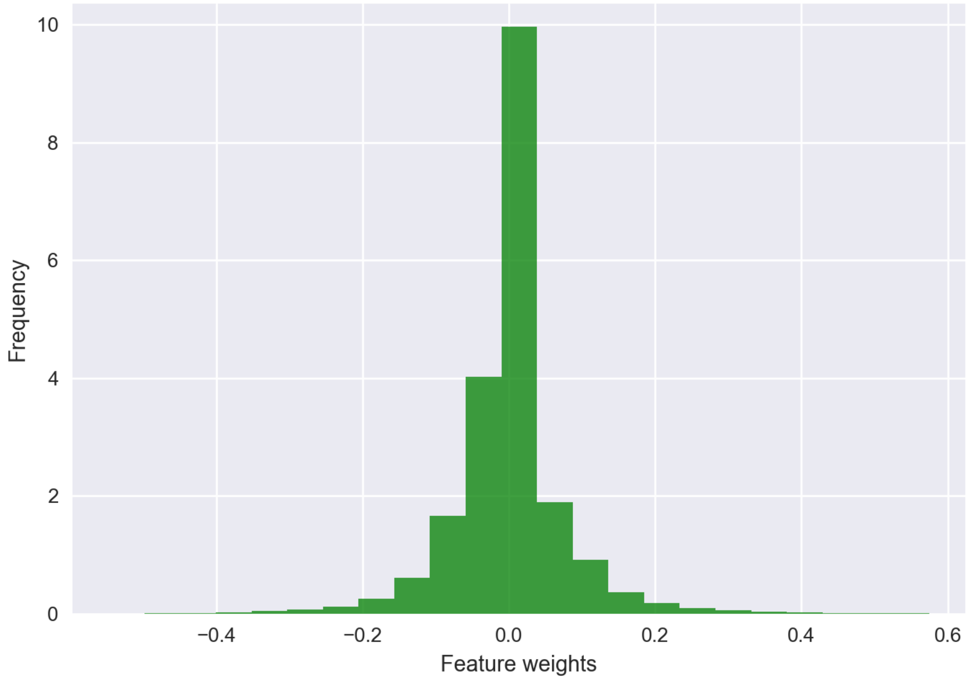

Figure 3. Histogram of the weights of the features learnt by the SVM classifier.

To gain further insights into the skip-gram patterns that are identified by the classifier to be useful for predicting whether there is a positive relationship between a drug and an event in a tweet, we plot the histogram of the feature weights in

Figure 3. From Figure 3., we see that the majority of patterns have their weights close to zero, and an almost identical spread in positive and negative directions centered around zero. We counted 60430 patterns to have weights exactly set to zero, meaning that approximately 35% (60430/168663) of patterns are found to be uninformative by the classifier. A randomly selected subset of zero-weighted patterns is shown in Table 1. We see that patterns that are likely to appear in tweets, irrespective of the tweet is about an ADR event are accurately pruned out by the classifier. Therefore, even if we have a comparatively larger feature space to the number of training instances, this does not necessarily result in overfitting.

Table 2: A randomly selected sample of features with zero weights.

Prefix patterns	Midfix patterns	Postfix patterns
P+trip+i	M+bad+idea	S+over
P+news+:	M+a+breakfast	S+12+hours
P+dat+lean	M+if+school	S+conquest
P+@rroddger	M+medica_authorities	S+please
P+fussiness+no	M+convicted+i	S+bad!

We list the top-ranked positively-weighted and negatively-weighted skip-gram patterns in Table 3. From Table 3 we see that skip-gram patterns that describe a positive relationship between a drug and an ADR are correctly identified by the proposed method. For example, the *P+took+too* indicates that the user has actually took the drug. Moreover, we see many negations in the top-ranked negatively-weighted patterns. Such clues could be used in several ways. First, we can use these clues as keywords for filtering social media posts that describe a potential positive relationship between drugs and ADRs. For example, we could run disproportionality-based signal detection methods using the disproportionality counts obtained from those filtered social media posts, thereby increasing the reliability of the detection. Second, these clues could be used to develop extraction patterns/templates that can be used for matching and extracting previously unknown ADRs for novel or existing drugs.

Table 3: Top-ranked positively (left two columns) and negatively (right two columns) weighted features (skip-gram patterns) by the SVM. P, M, S indicate respectively prefix, midfix, and postfix skip-gram patterns. For bigrams, we have used ’+’ to separate the constituent unigrams.

Feature	**weight**	**Feature**	**weight**
S+als	1.2096	M+commercial	-1.2304
M+induced	1.1314	P+hate+being	-1.0398
P+oh+no	1.0683	P+I’m+definitely	-1.0000
M+stinks	1.0000	P+clumsiness	-1.0000
S+.+wooh	1.0000	P+hospitalisation	-1.000

M+never+work	1.0000	S+lol+fml	-0.9674
P+high+off	0.9006	S+wopps	-0.9035
P+took+too	0.8449	P+rt+xanaaxhadme	-0.8067
M+was+supposed	0.8378	P+don't+think	-0.7721

Conclusions

We proposed a novel signal detection problem where given a social media post T that contains a drug D and an event R, we would like to determine whether R is related to D, or otherwise. We have then proposed a method to solve this signal detection problem utilising the lexical contextual information in T. Specifically, we extracted skip-gram patterns from the prefix, midfix, and suffix in T, and trained a binary SVM using a manually labelled training dataset. Our results show that the proposed method significantly outperformed the majority baseline and a bag-of-words classifier. Moreover, we showed that the discriminative patterns were ranked at the top by the trained classifier. In the future, we plan to use the automatically extracted patterns to develop an ADR extraction method for previously unknown adverse reactions of drugs from social media.

6. Acknowledgements

All authors of this paper were supported by the WEB-RADR projected funded by the Innovative Medicines Initiative (IMI).

7. References

[1] *Adverse Drug Reactions*, volume second edition. Pharmaceutical Press, 2006.

[2] I. Ahmed, F. Haramguru, A. Fourrier-Reglat, F. Thiessard, C. Kreft-Jias, G. Miremont-Salame, B. Begaud, and P. Tubert-Bitter. Bayesian pharmacovigillance signal detection methods revisited in a multiple comparison setting. *Statistical Methods*, pages 1774–1792, 28.

[3] Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Aritificial Intelligence Research*, 38:135–187, 2010.

[4] A. Bate, M. Lindquist, I.R. Edwards, S. Olsson, R. Orre, A. Lansner, and R.M. De Freitas. A bayesian neural network method for adverse drig reaction signal generation. *European Journal of Clininacl Pharmacology*, 54:315–321, 1998.

[5] A. Bate, M. Lindquist, I.R. Edwards, and R. Orre. A data mining approach for signal detection and analysis. *Drug Safety*, 25:393–397, 2002.

[6] D.W. Bates, R.S. Evans, H. Murff, P.D. Stetson, L. Pizziferri, and G. Hripcsak. Detecting adverse events using information technology. *Journal of the American Medical Informatics Association*, 10(2):115–128, 2003.

[7] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Measuring the similarity between implicit semantic relations from the web. In *WWW 2009*, pages 651 – 660, 2009.

[8] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. A relational model of semantic similarity between words using automatically extracted lexical pattern clusters from the web. In *Empirical Methods in Natural Language Processing*, pages 803 – 812, 2009.

[9] Steve Chaplin. The yellow card schele – why are gps under-reporting. *Journal of Prescribing and Medicines Management*, 17(15):18–22, 2006.

[10] Quang Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[11] W. DuMouchel and R. Harpaz. Regression-adjusted gps algorithm (rgps), 2012.

[12] W. DuMouchel and D. Pregibon. Empirical bayes screening for multi-item associations. In *Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 67–76, 20011.

[13] William DuMouchel. Bayesian data mining in large frequency tables, with an application to fda spontaneous reporting system. *The American Statistician*, 53(3):177 – 190, August 1999.

[14] David A. duVerle and Helmut Prendinger. A novel discourse parser based on support vector machine classification. In *ACL 2009*, pages 665–673, 2009.

[15] I.R. Edwards and J.K. Aronson. Adverse drug reactions: definitions, diagnosis, and management. *Lancet*, 356(9237):1255–1259, 2000.

[16] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[17] D.M. Fram, J.S. Alemenoff, and W. DuMouchel. Empirical bayesian data mining for discovering patterns in post-marketing drug safety. In *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 359–368, 2003.

[18] Miguel Angel Ríos Gaona, Alexander Gelbukh, and Sivaji Bandyopadhay. Recognizing textual entailment with statistical methods. In *Proc. of the Mexican Conference on Pattern Recognition*, pages 372–381, 2010.

[19] K.M. Giacomini, R.M. Krauss, D.M. Roden, M. Eichelbaum, M.R. Hayden, and Y. Nakamura. When good drugs go bad. *Nature*, 446(7139):975–977, 2007.

[20] Manfred Hauben and Xiaofeng Zhou. Quantitative methods in pharmacovigilance. *Drug Safety*, 26(3):159 – 186, 2003.

[21] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jiang Yang, and Graciela Gonzalez. Towards internet-age phramacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. In *Proc. of ACL 2010 Workshop on Biomedical Natural Language Processing*, pages 117 – 125, 2010.

[22] William C Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243 – 281, 1988.

[23] Daniel Marcu and Abdessmad Echihabi. An unsupervised approach to recognizing discourse relations. In *ACL'02*, pages 368 – 375, 2002.

[24] G.N. Noren, A. Bate, R. Orre, and I.R. Edwards. Extending the methods used to screen the who drug safety database towards analysis of complex association and improved accuracy for rare events. *Statistical Methods*, 25:3740–3757, 2006.

[25] Diarmuid Ó Séaghdha and Simone Teufel. Unsupervised learning of rhetorical structure with un-topic models. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2–13, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[26] World Health Organisation. International drug monitoring: The role of the hospital. 1966.

[27] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *WWW'12*, pages 909 – 918, 2012.

[28] Majid Rastergar-Mojarad, Hongfang Liu, and Priya Nambisan. Using social media data to identify potential candidates for drug repurposing: A feasibility study. *JMIR Research Protocols*, 5(2), 2016.

[29] Isabel Segura-Bedmar, Paloma Martinez, and Maria Segura-Bedmar. Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems. *Drug Discovery Today*, 13(17–18):816–823, September 2008.

[30] Richard Sloane, Orod Osanlou, David Lewis, Danushka Bollegala, Simon Maskell, and Munir Pirmohamed. Social media and pharmacovigilance: A review of the opportunities and challenges. *British Journal of Clinical Pharmacology*, 80(4):910–920, 2015.

[31] M. Suling and I. Pigeot. Signal detection and monitoring based on longitudinal healthcare data. *Pharmaceutics*, 4:607–640, 2012.

[32] Cornelis S. van Der Hooft, C.J.M. Miriam, Kes van Grootheest, Herre J. Kingma, and H. Brino. Adverse drug reaction-related hospitalisations: a nationawide wide study in the netherlands. *Drug Safety*, 29(2):161–168, 2006.

[33] V. Vapnik. *Statistical Learning Theory*. Wiley, Chichester, GB, 1998.

[34] Christopher C. Yang, Haodong Yang, Ling Jiang, and Mi Zhang. Social media mining for drug safety signal detection. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, SHB '12, pages 33–40, New York, NY, USA, 2012. ACM.

[35] MedWatch, https://yellowcard.mhra.gov.uk/

[36] YellowCard System, http://www.fda.gov/Safety/MedWatch/

[37] Twitter, www.twitter.com

[38] Facebook, www.facebook.com

[39] Pinterest, www.pinterest.com

[40] Instgram, www.instgram.com

[41] WEB-RADR, https://web-radr.eu

[42] MedDRA, www.meddra.org

[43] Twitter Streaming API, https://dev.twitter.com/streaming/overview

[44] MongoDB, www.mongodb.com

[45] Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. Pharmacov- iligance from social media: mining adverse drug reaction mentions

using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association, 22(3):671–681, 2015.

[46] Jeremy Lardon, Redhouane Abdellaoui and Florelle Bellet, Hadyl Asfari, Julien Souvignet and Natalie Texier, Marie-Christine Jaulent and Marie-Noelle Beyens and Anita Burgun and Cedric Bousquet. Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review. Journal of Medical Internet Research, 17(7), e171, July, 2017.

[47] Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. Adverse drug reaction classification with deep neural networks. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 877–887, December 2016.

[48] Yoon Kim. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Confer- ence on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, 2014.

[49] Kathy Lee, Ashequal Qadir, Sadid A. Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In Proc. of International World Wide Web Conference, pages 705–714, 2017.

[50] Anne Cocos, Alexander G Fiks, and Aaron J Masino. Deep learning for pharamacovigilance: recurrent neural network architectures for labelling adverse drug reactions in twitter posts. Journal of the American Medical Informatics Association, 24(4):813–821, July 2017.

[51] Yoon Kim. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Confer- ence on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[52] A. Culotta and J. Sorensen. Dependency tree kernels for relation extraction. In Proc. of ACL'04, pages 423–429, 2004.

[53] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In 5th ACM International Conference on Digital Libraries, pages 85–94, 2000.

[54] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, 26(4):404, 1934.

[55] Nut Limsopatham and Nigel Collier. Adapting phrase-based machine translation to normalise medical terms in social media messages. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1675–1680, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[56] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12:2121 – 2159, July 2011.