Neural Processing Letters (2020) 52:1703–1721
https://doi.org/10.1007/s11063-020-10296-7

Feature-Based Learning in Drug Prescription System for Medical Clinics

Wee Pheng Goh[1] · **Xiaohui Tao**[1] · **Ji Zhang**[1] · **Jianming Yong**[1]

Published online: 2 July 2020

Abstract
Rapid increases in data volume and variety pose a challenge to safe drug prescription for health professionals like doctors and dentists. This is addressed by our study, which presents innovative approaches in mining data from drug corpus and extracting feature vectors to combine this knowledge with individual patient medical profiles. Within our three-tiered framework—the prediction layer, the knowledge layer and the presentation layer—we describe multiple approaches in computing similarity ratios from the feature vectors, illustrated with an example of applying the framework in a typical medical clinic. Experimental evaluation shows that the word embedding model performs better than the adverse network model, with a F score of 0.75. The F score is a common metrics used for evaluating the performance of classification algorithms. Similarity to a drug the patient is allergic to or is taking are important considerations for the suitability of a drug for prescription. Hence, such an approach, when integrated within the clinical work-flow, will reduce prescription errors thereby increasing patient health outcomes.

Keywords Feature vector · Similarity ratio · Word embedding · Adverse network model · Personalised drug prescription

This work is partially supported by Glory Dental Surgery Pte Ltd and undertaken collaboratively with their panel of dentists. We would like to thank Dr. Xueling Oh and Dr. Elizabeth Goh for enriching the authors understanding of drug prescription within the medical domain.

✉ Wee Pheng Goh
weepheng.goh@usq.edu.au

Xiaohui Tao
xtao@usq.edu.au

Ji Zhang
ji.zhang@usq.edu.au

Jianming Yong
jianming.yong@usq.edu.au

[1] University of Southern Queensland, Toowoomba, Australia

1 Introduction

The increasing amount of data available to medical professionals for diagnosing diseases and developing treatment plans for their patients raise the importance of having suitable tools to harness such data and transform them into meaningful information. Such tools are also useful for evidence-based decision making within the healthcare domain [14]. Health-care professionals can no longer solely rely on pen and paper as more and more data are in digital form. The skill to click, copy and paste is becoming more crucial than the ability to hold a pen, flip and clip papers. X-ray films are replaced by digital X-ray. Treatment notes written on cards are replaced by digital notes. Lead used in pencil is progressively replaced by silicon used in electronic devices for writing and drawing notes. Wooden furniture for storage is being replaced by digital media. Hence a decision support system is important for the health practitioner to deliver the service efficiently [10].

In terms of drug prescription, not only more information is to be stored and retrieved from digital media, the number of drugs that doctors need to handle is also increasing as more and more patients are taking multiple drugs. Within dental clinics, antibiotics are often used to resolve infections especially after surgical procedures such as placement of dental implants and gum treatment [26]. To manage anxiety which usually occurs before surgery [9], anxiolytic drugs are also commonly used. Reducing and relieving pain with analgesic medications are also important procedures within the clinical work-flow of any health institution [22].

Although there are many studies that examine drug–drug interactions (DDI) [3,4,32], they do not associate them with the patients medical profile to facilitate individual drug prescription. Although our system is similar to that proposed by Casillas et al. [5] in terms of using information from the patient, the unique approach adopted in this paper goes one step further in using such information to support the decision-making process for doctors at point-of-care within the clinical work-flow. An additional presentation layer is introduced, providing an important interface between the user and the knowledge mined from bio-medical sources. In addition, within the predictive layer, we propose multiple approaches to deduce the similarity ratio between drugs in a drug-pair to assist doctors in prescribing drugs at point-of-care, in order to find the combination of approaches that yields the best performance. This is in contrast with only a single network model approach in our previous work [11].

Knowledge obtained from data mining performed on open source datasets to predict the relationship between drugs in a drug-pair is used to give prescription support to the health practitioner. Feature vectors will be built from the text corpus to allow computing of the similarity ratio, with the assumption that a drug-pair safe for consumption will have a higher similarity ratio compared to an unsafe pair. These multiple methods for building the feature vectors reside within the prediction layer of our three-tier framework.

Experimental results show that the word-embedding model performs better than the adverse network model. Performance is also better than the baseline model. This model is easily utilised in predicting a drug's suitability for prescription by considering the patients drug allergies to avoid allergic reactions, and the drugs the patient is currently taking to avoid adverse DDI.

This study will help provide strategies in research agenda and priorities, including methodologies for knowledge reasoning and inference in the context of a medical clinic. Research outcomes of this project, especially in this climate of increasing poly-pharmacy, will help reduce the risk of prescribing drugs that may cause the patient to suffer an adverse reaction and thus improve healthcare quality. This system which delivers information on interacting

drug-pairs based on the patient's drug profile will also benefit those who are involved in clinical education and research relating to drug dispensing, such as medicine, nursing and pharmacy.

Traditionally, chemical structures and drug targets were used to decide if a drug-pair is interacting. By using data mining and feature extractions from text corpus, this paper contributes significantly in the way useful information on drugs interactions can be obtained. Moreover, we will also show in this study how such information can be applied for personalised decision support in the area of drug prescription within a medical clinic.

The efficient approach in the design of the clinical decision support system (CDSS) with consideration of the medical profile of the patients result in the following significant contributions:

- advancement in the design of clinical decision support systems by using similarity ratio of a drug-pair;
- attributes like adverse interactions and side effect of a drug can be used to construct feature vectors for computing similarity ratios;
- by hierarchically representing the drug-pairs within the context of a CDSS, paths linking the common drugs within the set of interacting drugs can be used to arrive at a similarity ratio;
- results support the hypothesis that similar drug-pairs have a higher similarity ratio compared to that of dissimilar pairs;
- provide a platform for further research on data mining and machine learning methods within the medical domain which will transform the clinical work flow of the health-care industry.

The rest of the paper is organised as follows: Sect. 2 discusses the related work in data mining and how our model differs in the way the drug–drug relationship is detected and deployed for use. The multi-model framework is described in Sect. 3, while Sect. 4 describes the experiment and Sect. 4.6 discusses the results, with a description on how the model can be applied in a medical clinic. Finally Sect. 5 presents the conclusions obtained.

2 Related Work

Word embedding is a method to represent the semantic and syntactic similarities between words. It has found application in many areas including sentiment analysis [27] and sentence classification [33]. Inspired by deep neural network models, word embeddings have drawn the interest and attention of many researchers.

The ability to predict context words has motivated many studies to use this model for obtaining the similarity of drugs within drug-pairs. A recent work by Wang et al. [30] used this approach to extract information on DDI from biomedical corpus, by capturing the core meaning of the sentences in the text and incorporating the syntactic contexts into the embeddings. [36] examined the ability of word2vec in deriving semantic relatedness and similarity between biomedical terms in journal articles. It is interesting to note that models trained on specific text like abstracts yielded better results that those trained with the main text of the articles. [17] attempted to use word embeddings to capture semantic information of words for the DDI classification. [35] also used word embeddings to exploit the syntactic information of a sentence to extract DDI. Both systems delivered promising performance, despite neither being customised to the patient's individual drug profile.

There has been growing interest in comparing text and computing similarity between entities by representing them in a graphical model. For example, in Palma et al.'s model, the semantics similarity between drugs is used to predict drug target interactions [20]. Based on the hypothesis that similar targets interact with the same drugs, and similar drugs interact with the same targets, a heterogeneous graph was constructed with edges that include the drug–target interaction as well as drug–drug and target–target similarity edges.

[12] also proposed a framework to compute the similarity between two objects by representing them and their relationship as a graph. With the objects as nodes and their relationship as edges, this framework assumes that two objects are similar if the objects related to them are also similar. For example, two publications are considered similar if the papers cited by each publication are also similar. The directed graph $\mathcal{G}$ used to represent such a framework with nodes V and edges E can be formally defined as G = (V, E) where the nodes V represent the objects and the edges E represent the relationship between the objects. If $I_i(v)$ represents individual incoming objects and $O_i(v)$ individual outgoing objects, then the similarity score between any two nodes A and B is given by:

$$s(A, B) = \frac{C1}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} s(O_i(A), O_j(B)) \tag{1}$$

In another model with special emphasis on Heterogeneous Information Networks (HIN) [29], similarities between two entities can be found by considering the number of paths between them. Nodes and edges are defined in this HIN as $G = (V, E)$ where the nodes are the set of entities A and edges are the set of links R between the items in the entities. The entity type mapping is given by $\phi = V \rightarrow A$ and the relation type mapping given by $\psi = E \rightarrow R$. The similarity between two entities A and B is defined as:

$$\text{s(A,B)} = \frac{\text{2.\# paths between A and B}}{\text{\#circles with entity A} + \text{\#circles with entity B}} \tag{2}$$

In yet another attempt to represent entities in a directed graph, Shi et al. focused on an approach which also assigns weights to the relations between the entities [25]. Hence, this method of representation becomes appropriate for a recommender system. Besides weightage, this method also assigned attributes to the links between the entities. For example, users a and b may have a common liking for movie $m1$ (Fig. 1), as well as other movies, so we can say that $m1$ is in a direct neighbourhood of a.

Although there has been much research on DDI using different techniques, there is no system that uses DDI information to facilitate drug prescription within a CDSS, notwithstanding the absence of a complete source of information on potential DDI [1]. A CDSS that conforms to our recommendations of a personalised system, which considers the drugs the patient is taking and is allergic to, will contribute to the productivity and efficiency of medical treatment, with practitioners more readily adopting such a system within their clinical work-flow [10]. Therefore, a CDSS which integrates with drug knowledge bases to identify adverse drug events and advises on drug suitability before prescription will appear helpful to the health practitioner. With timely and accurate DDI information embedded within a CDSS, more comprehensive treatment options can be made available to patients and practitioners, thus contributing to a more positive treatment experience, better oral health outcomes and job satisfaction for the medical practitioner.

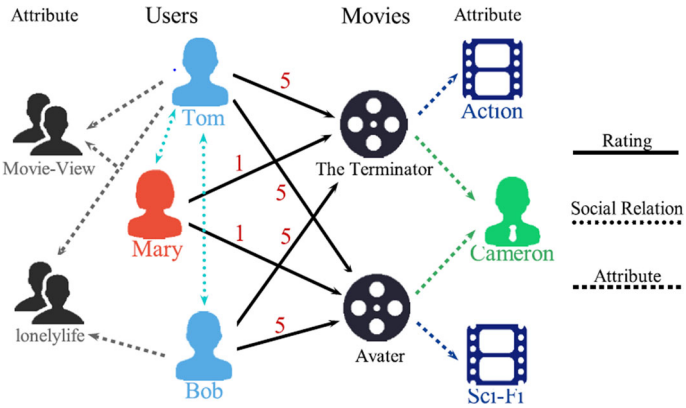

Fig. 1 Objects and relations in an information network, taken from [25]

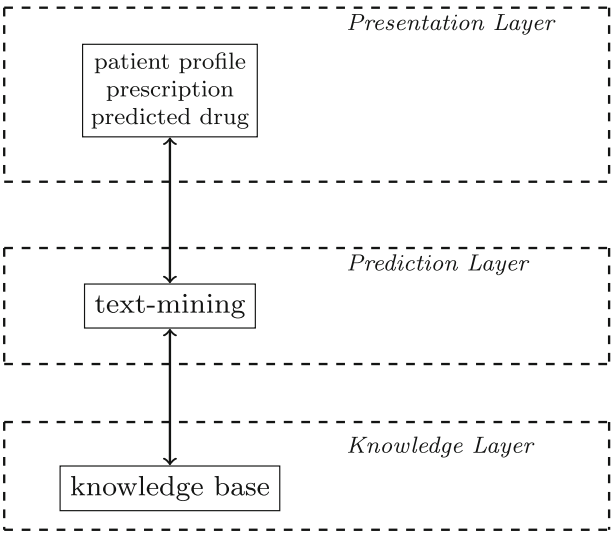

Fig. 2 Three-tier framework

3 The Proposed Framework

A three-tier framework was proposed to allow ease of design and portability across different fields in medical diagnosis and prescription support (Fig. 2). This framework consisted of the knowledge layer, the prediction layer and the presentation layer. Such a framework allowed each layer to be developed and maintained independently, while at the same time ensuring the inter-layer interfacing conformed to standards.

For example, by using the word embedding approach in building feature vectors, the output from the prediction layer can be fed into other neural networks to accomplish other tasks. In addition, this framework supports our unique approach of personalised drug prescription. The following sections describe the functions of each layer.

Table 1 Co-occurrence matrix

	Scientist	Research	Risk	Factor	Covid-19
Scientist	0	0	0	0	0
Research	0	0	1	1	0
Risk	0	1	0	1	0
Factor	0	1	1	0	1
Covid-19	0	0	0	1	0

3.1 Knowledge Layer

The knowledge layer consists of the biomedical text which describes the properties of the drugs. The text comprises of a bag of words from which relevant information was extracted at the data mining layer for computing similarity ratio within a drug-pair.

In our study, the text from DrugBank was used as it has the advantage of having each drug described with different properties from different perspectives. It contains a comprehensive corpus of information to suit both patients (under the heading "Overview" and healthcare professionals (under the heading "Professionals" while information on side-effects are found under the heading "Side-effects". An updated knowledge base is also important for the system to be perceived as useful and adopted by the user [13]. All this information is collectively stored in the drug taxonomy $\mathcal{T}$, defined as a 3-tuple $\mathcal{T} := \langle \mathbb{D}, \mathbb{R}, \mathcal{H}_{\mathbb{D}}^{\mathbb{R}} \rangle$, where

- $\mathbb{D} = \{d_1, d_2, \ldots, d_{|\mathbb{D}|}\}$ is the domain set of drugs;
- $\mathbb{R} = \{r^+, r^-, r^0\}$ is a set of semantic relations, where $r^+(d_i, d_j)$ means that the effects of drugs d_i and d_j are advantageous; $r^-(d_i, d_j)$ means that the effects of drugs d_i and d_j are adverse; $r^0(d_i, d_j)$ means that the effects of drugs d_i and d_j are not related.
- $\mathcal{H}_{\mathbb{D}}^{\mathbb{R}}$ is the taxonomical structure constructed by all $d \in \mathbb{D}$ linked by $r \in \mathbb{R}$.

This layer contains important information relating to DDI and provides the ground truth in deciding if a drug-pair has an adverse relationship.

In order to construct the knowledge base for use by the models during the experiment, textual data that describes each drug within drugBank is extracted by parsing through drugBank from the different properties for use in the knowledge base. The collection of these drug properties then goes through the pre-processing stage. This is to ensure that non-paltry terms are not omitted in the subsequent assembling of patterns of words for building features of the drugs. At this stage, stopwords were removed and words converted to their root form through stemming which enhanced the reliability of the data [23]. With these words being collected, patterns of each word can be constructed to allow similarity ratio to be computed in the models residing in the Prediction Layer during the experiment.

Take for example the sentence "Scientists are still researching risk factors for COVID-19" found in the drugBank repository of documents. After going through the process of removing stopwords and transforming appropriate words to their root form, the words in the sentence are reduced to "scientist","research", "risk", "factor" and "COVID-19".

A co-occurrence matrix with these keywords can be constructed so that each word can be represented as a pattern of binary digits. The placement of the binary digits depends on the rule dictating the contextual distance before and after each word. Hence Table 1 shows the co-occurrence matrix which indicates the occurrence of the words together within a contextual distance of two words before and after each target word. It can be seen that for the target word "factor", within a context distance of two, each of these words "research", "risk" and "covid-19" occurred once as indicated in the row beginning with "factor" in Table 1. Thus,

Table 2 Features of conceptual framework

Presentation layer	Prediction layer	Knowledge layer
• Efficient mapping of user requirements	• Efficient choice of programming approach	• Bio medical data data sources, drug taxonomy
• User-friendly interface	• Implementation of data mining	• Drug properties
	• Algorithm design	

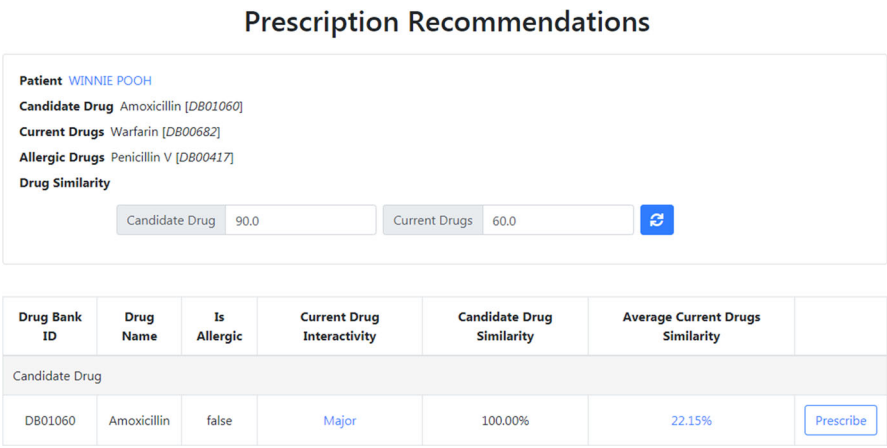

Fig. 3 User interface

the row matrix for "factor" can be represented as $\begin{bmatrix} 0 & 1 & 1 & 0 & 1 \end{bmatrix}$. With each word in the textual description of the drugs in the bio-medical database being represented as a co-occurrence matrix, a knowledge base associated with each drug can be constructed. This will facilitate subsequent building of the drug model to determine their similarity ratio.

3.2 Presentation Layer

The presentation layer is important as it serves as an interface between the computing layer and the user. A well-designed user-friendly interface will help users adopt such a system in their clinical work-flow. As highlighted in Table 2 for the three layers in the framework, user requirements in the presentation layer need to be efficiently mapped onto the prediction layer to enable useful and relevant information to be extracted for further computing of the similarity ratio.

The presentation layer also distinguishes our system from many other decision support systems as it contains the patient's personalized information.

In this system, patient p was defined as a 2-tuple $p := \langle \mathcal{D}, \mathcal{D}^- \rangle$, where

- $\mathcal{D} \subset \mathbb{D}$ is the set of drugs that p is currently taking, where $|\mathcal{D}| <= \theta_{\mathcal{D}}$;
- $\mathcal{D}^- \subset \mathbb{D}$ is a known set of drugs that p is allergy to, where $|\mathcal{D}^-| <= \theta_{\mathcal{D}^-}$ and $\mathcal{D}^- \cap \mathcal{D} = \emptyset$.

Besides, the presentation layer also presents the results from the prediction layer. Hence this layer is important as a supporting tool to doctors in deciding whether the drug to be prescribed is safe for the patient.

The drug to be prescribed is also stored in this layer. Such information is needed in the prediction layer for extraction of feature vectors. In order to maintain user-friendliness, which is crucial for clinical adoption of the system, it is important for this layer to present the results in a user-friendly manner.

Based on the results transmitted from the prediction layer, the service at this layer will then advise the user if the drug in question is safe for prescription. This approach allows the presentation layer to crystallise the results in a meaningful and friendly manner. This will allow a prescription similar to Fig. 3 to be presented to the user.

3.3 Prediction Layer

From the drug taxonomy $\mathcal{T}$, text for each drug was extracted, cleaned and stored in order to provide information on the underlying properties of a drug-pair. The flexibility and robustness of the three-tier framework allowed the calculation of drug-pair similarity using various approaches.

As individual drugs and the adverse relationship between them can be logically represented in a network of nodes and edges, the adverse network model is used in computing the similarity between drugs within a drug-pair. In fact, such an information network allows rich structure and semantic information to be stored which enables further research associated with data mining [24].

Besides the adverse network model, the word embedding approach is also used to compute the similarity ratio of a drug-pair. This method of finding the similarity between a drug pair is adopted due to its increasing popularity in machine learning. In tasks involving word similarity, recent trends also suggested the use of word embedding models as they outperform other traditional models like the count-based distributional models [16].

3.3.1 Adverse Network Model

In the adverse network model, drugs and their relationships are represented by a graph $G = (V, E)$. Basically, we adapt Jeh et al.'s model of measuring similarity based on theoretical foundations [12]. In our model, we represent all drugs as nodes in a network to enable us to compute their proximity in terms of the number of shared entities between the drug pair.

For a given node A, $O(A)$ denotes the set of out-neighbors and the number of out-neighbors of node A is $|O(A)|$. Similarity between node A and node B will be given by:

$$s(A, B) = \frac{C1}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} s(O_i(A), O_j(B)) \tag{3}$$

Suppose the set of interactive drugs is defined as $A^r = \{a_1, a_2 \ldots a_k\}$, where r is the attribute of the relationship with the vertex drug. The items in A were also the subset of out-neighbours of drug d_1, denoted as $O(d_1)$. Individual drugs in A which interact with d_1 were then denoted by $O_i(d_1)(1 \leq i \leq |O(d_1)|)$. Hence, $O_i(d_1) \in \{A^r\}$.

Referring to Fig. 4, if we consider the out-neighbours, drug v_1 will have a set of interacting drugs $O(v_1)$, where the number of drugs $a_1, a_2, a_3 \ldots a_k$ that adversely interacts with drug v_1 is $k = |O(v_1)|$.

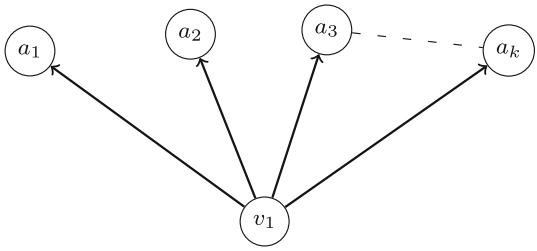

Fig. 4 Graph of interactive drugs with drug v_1

Path connecting two nodes indicates the relationship between the two drugs in a drug-pair. Adapting [25]'s notation, the path from drug d_1 to the set of interactive drugs A with ratings r is denoted by $d_1 \xrightarrow{r} A$, which can also be written as $d_1(r)A$, where r is the relationship between d_1 and A. Thus, $d_1(1)A$ shows drug d_1 has minor interaction with the drugs in set A.

3.3.2 Word Embedding Model

The second framework used the skip-gram model, which is commonly utilised for learning word embeddings by predicting context words given a target word as the input to the model. With the context words, feature vectors were then extracted through word embeddings, which transform the words into vectors. Since it is expected that a larger set of common terms is used to describe a pair of drugs that are similar in function, it follows that the similarity between drugs in a drug-pair can be measured by finding words that are most related to each drug in the drug-pair.

Such an approach to machine learning has already made major impact in many areas such as medical imaging, speech recognition and natural language processing where a large amount of data is involved, and is very relevant considering the constant increase of drug-related biomedical information [31]. Interest in word embedding has also resulted in studies on the influence of domain type and size on their performance [15].

One of the reasons for its popularity lies in the fact that analogical linguistic relationships among words can be easily discovered through word embedding. Interest in word embedding has intensified with Mikolov et al.s introduction of a simplified architecture, which eliminates the non-linear hidden layer, allowing training on much larger datasets than was previously possible [18].

Instead of using the set of interactive drugs as in the previous model, data from the text corpus was used in this model to compute the similarity within a drug-pair. With the help of Word2Vec, tokens were then built by iterating through the sentences in the text corpus, specifying parameters such as minimum word frequencies and the size of the feature vectors. Word2Vec is used as it has been reported to be the most efficient ones for learning vector representations of words [19]. While Word2Vec is not strictly a deep neural network, the output vector that it produces in numerical format within the deep learning models can be easily understood by other deep networks making it very suitable for use in such works.

Assuming w_c is the context word and w_t is the target word, the goal of the skip-gram model is to maximise the log-likelihood of obtaining the output context word given the input target word, ie,.

$$J = log\, P(w_c|w_t) \tag{4}$$

where J is the objective function.

Suppose u_{wt} is a target embedding vector for w_t and v_{wc} is a context embedding vector for the context word w_c, then P, the conditional probability in the neural probabilistic language model can be defined as:

$$P(w_c|w_t) = \frac{exp(v_{wc}^T u_{wt})}{\sum_{w=1}^{W} exp(v_{wc}^T u_{wt})} \tag{5}$$

Taking the log on both sides of Eq. 5 above,

$$log P(w_c|w_t) = log \frac{exp(v_{wc}^T u_{wt})}{\sum_{w=1}^{W} exp(v_{wc}^T u_{wt})} \tag{6}$$

Since J = log P (from Eqs. 4), 6 becomes:

$$J = logexp(v_{wc}^T u_{wt}) - log \sum_{w=1}^{W} exp(v_{wc}^T u_{wt}) \tag{7}$$

In order to avoid expensive computation of softmax for the whole vocabulary, negative sampling is commonly used. Then the objective function J becomes:

$$J' = \sum_{w_t, w_c \in D} log\, Q_\theta(D = 1|w_t, w_c) + \sum_{w_t, w_c \in D'} log\, Q_\theta(D = 0|w_t, w_c) \tag{8}$$

With the probability of w_t and w_c being observed is $Q_\theta(D = 1|w_t, w_c)$ and the probability of not being observed is $Q_\theta(D = 0|w_t, w_c)$, D and D' is the observed data and unobserved data respectively and θ word embeddings.

Once the text corpus has been trained by Word2Vec, the output vector for any name of a drug can be conveniently obtained through built-in Java methods included in Word2Vec. For example, given a keyword, the output vector comes in an array of numbers, and the size of such arrays depends on the number of nearest neighbours specified in the experiment. The more frequently the combination of words occurred in the training sample, the more likely the word would be selected. The layer size determined the size of the output feature vectors. Thus, if the vocabulary size of the corpus was k, and the number of terms in the text corpus was n, the input vector would be a single row vector [1 x k] containing 0 at all positions within the vector except the nth position which would be a 1. With a layer size of m, the size of the hidden layer used by Word2Vec is [k x m]. In this way, the word vector produced for each word would be the product of the matrix [1 x k] and [k x m] producing a single row vector of size m.

4 Experimental Evaluation

The performance of our novel approach was evaluated by individual testing of each model, as well as combined testing in a sensitivity study. The same training set used for all the experiments consisted of positive and negative drug-pairs according to the drug taxonomy. The positive set contained drug-pairs that do not adversely interact with one another whereas the negative set contained pairs which adversely react with one another and should not be prescribed together. A similarity ratio above a threshold value of 0.5 implies the model is making a correct prediction.

4.1 Similarity Ratio

The similarity ratio used in the experiment is based on the cosine similarity between two feature vectors. If feature vectors of drug d_i and drug d_j are given by:

$$\begin{aligned} \overrightarrow{d_i} &= \{a_1, a_2, a_3 \ldots a_n\} \\ \overrightarrow{d_j} &= \{b_1, b_2, b_3 \ldots b_n\} \end{aligned} \tag{9}$$

where $a_1, a_2 \ldots a_n$ are the vector components of drug d_i and $b_1, b_2 \ldots b_n$ are the vector components of drug d_j.

Their dot product will be

$$\overrightarrow{d_i} . \overrightarrow{d_j} = \sum_{k=1}^{n} a_k b_k = a_1 b_1 + a_2 b_2 + \ldots a_n b_n \tag{10}$$

and the geometric definition is given by

$$\overrightarrow{d_i} . \overrightarrow{d_j} = \| \overrightarrow{a} \| \| \overrightarrow{b} \| cos\theta \tag{11}$$

Rearranging gives

$$cos\theta = \frac{\overrightarrow{d_i} . \overrightarrow{d_j}}{\| \overrightarrow{a} \| \| \overrightarrow{b} \|} \tag{12}$$

The angle θ represents the similarity between two documents represented by vectors $\overrightarrow{d_i}$ and $\overrightarrow{d_j}$. Depending on the model used during the experiment, various ways are used to obtain the feature vectors $\overrightarrow{d_i}$ and $\overrightarrow{d_j}$. If both documents contain similar terms, their feature vectors will be similar. In other words, the angle θ_1 between the two vectors in the vector space will be small, as both are heading closely in the same direction.

Conversely, if the drug-pair d_i and d_j contains more dis-similar terms, then the vectors will be heading at a larger angle resulting in a smaller cosine similarity since cos θ_2 is lesser than cos θ_1 when θ_2 is larger than θ_1. In fact, if there are totally no common terms, the two vectors are said to be perpendicular to each other or orthogonal which result in zero similarity ratio since cos 90° is zero.

4.2 Experimental Design

The training set consisted of sample drug-pairs that were either similar (true positive) or dissimilar (true negative) according to the drug taxonomy.

4.2.1 Adverse Network Model

During the experiment, matrix M^r for drugs d_1 and d_2 was created to indicate the positional match in the adverse drugs for d_1 and d_2 with interaction rating r where number of columns in $M^r = |A^r|$.

If $O_i(d_1)$ is found in A^r at position u, then the uth column in matrix M^r will be updated as $r, M^r(1, u) \longleftarrow r$. For example, Table 3 shows the row matrices for the case of an interaction between drug pair v_1 and v_2 at a rating of 3. Adverse drug with v_1 is found at position $u = 5$ and adverse drug with v_2 is found at position 4 and position 5.

Table 3 Row matrix M^r at r = 3

Column u	1	2	3	4	5	6
Drug v_1					3	
Drug v_2				3	3	

The vectors obtained for drug v_1 and v_2 are $\mathbf{F_1^r} = \{a_1, a_2 \ldots a_p\}$ and $\mathbf{F_2^r} = \{b_1, b_2 \ldots b_p\}$ respectively where

$$a_u = \begin{cases} r, \text{ if } O_i(v_1) == A^r[u] \\ 0, \text{ otherwise} \end{cases}$$
$$b_u = \begin{cases} r, \text{ if } O_i(v_2) == A^r[u] \\ 0, \text{ otherwise} \end{cases} \tag{13}$$

Hence the similarity ratio between drug v_1 and drug v_2 can be obtained:

$$S^r(v_1, v_2) = \frac{\sum_{i=1}^{p} a_i \times b_i}{\sqrt{\sum_{i=1}^{p} a_i^2} \times \sqrt{\sum_{i=1}^{p} b_i^2}} \tag{14}$$

With the similarity ratio results, a threshold of $\theta = 0.5$ was used to predict if the drug-pair is similar. A value of 0.5 or higher from the experiment meant the drug-pair was considered similar, while a value below 0.5 meant the drug-pair was considered dissimilar. The models performance can be measured by counting the number of correct predictions.

4.2.2 Word Embedding Model

Since the aim is to discover the similarity between two drugs, it would be interesting to explore alternative measures in building feature vectors. In this system, feature vectors were obtained through an artificial neural network approach. By using the skip-gram model from Word2Vec [18], a predictive model was constructed for learning word-embeddings from the raw corpus that described the properties of the drugs. Since the problem domain aims to extract related words to determine the extent of similarity from biomedical text, word2Vec was relevant to our experiment. Given a keyword, for example, the drug name, this method formulated a feature vector that best predicts a window of surrounding words that occur in some meaningful context. Such semantic similarity also conforms to the important criteria for selecting good word pairs ([34])

When training the dataset, the parameters required by word2Vec were the word frequency (the minimum number of times a word must appear in the corpus), layer size (the number of desired features in the word vector) and window size (the number of words before and after the word to extract for the training sample).

With this model, word vectors were constructed by sending a keyword. During the experiment, keywords associated with the nearest neighbour of the drug name were retrieved from the model. Similarity ratio between each set of vectors produced from the keywords could then be computed. To observe the behavior of this approach, the model was constructed with individual properties of the drug ("Overview","Professional" and "Side Effects") while varying the number of nearest neighbours.

With each model, $word2vec(Ov)$, $word2vec(Pr)$ and $word2vec(Se)$ trained from the text corpus "Overview", "Professional" and "Side Effects" respectively, various keywords

Table 4 Size of dataset for used for building word2Vec model

Overview	Professional	Side-effects
154,645	196,352	53,644

could be obtained. Hence, each keyword was represented by a word vector of numbers, the size of which depended on the layer size as explained in Sect. 3.

In each model, word vectors were constructed from different combinations of keywords associated with the drug name. For example, if d_{11}, d_{12}, d_{13} were the three nearest keywords for a given drug d1, a word vector would be obtained from the specified model by combining the three word vectors from the respective three keywords.

4.3 Data Preparation

Drug pairs used in the experiment were extracted from DrugBank, an unique resource containing a comprehensive corpus of information relating to various properties of drugs relevant to both end-users and professionals. It is maintained in collaboration with the US Food and Drug Administration (FDA). This corpus contains 6811 drug entries including 1528 FDA-approved small molecule drugs, providing free, independent, peer-reviewed, and up-to-date information at both consumer and professional levels.

In order to prepare data for the experiment, textual data from DrugBank is downloaded and cleaned by removing stopwords with words converted to their root form through stemming. Table 4 shows the number of tokens for each attribute of the drug used for the experiment. For the word embedding model, these tokens are further used to build the binary model to be used for the experiment.

4.4 Baseline Model

Our work was evaluated against other works to highlight how adoption of this novel approach results in superior performance. The work of [28] predicted DDI by parsing biomedical text for syntactic and semantic information on biological entities such as induction and inhibition of enzymes by drugs. These relations were then mapped with the general knowledge about drug metabolism and interactions to derive the DDI. The work by [32] developed various prediction models to leverage on text mining and statistical inference techniques.

One of the models used include the popular DET model used to capture the relation between drugs and other entities. Using plate notations [2] Fig. 5 shows the generative process represented as a Bayesian model. A dummy document with subject section and content section is built for each drug found in the Medline corpus, assuming total number of drugs is D drugs. Hence total number of documents is D, with each document d conveyed by diseases. N_d refers to the total number of disease words w occuring in d. Total number of topics is K and a_d is the observed set of drugs, with A referring to the total number of drugs.

Each drug x and topic z is a probabilistic distribution over topics (parameterised by θ) and diseases (parameterised by φ) respectively. λ is the observable parameter which controls the drugs sampling.

Just like our work, DrugBank was also used. However, one of the methods in their preparation of data was to represent each drug by a vector of drug targets. The values in each vector are either 1 or 0, depending on whether the drug target is associated with the given drug. In

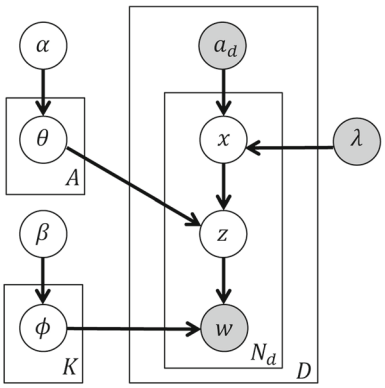

Fig. 5 Drug-entity model ([32])

our work, we chose to construct feature vectors of *tf*idf* from textual information related to the properties of each drug.

4.5 Performance Measuring Schemes

Precision, recall and F-measure were used to evaluate the performance of our model. Precision indicated how accurately the model predicted drug-pairs as similar, while recall indicated how accurately similar drug-pairs were predicted. Accuracy was also used to measure the percentage of correct predictions combining both the similar and dissimilar predictions.

4.6 Results and Discussions

With the unique three-tier conceptual framework where knowledge is extracted from the knowledge base and delivered to the prescription layer, the ensuing results demonstrate our model's efficiency and robustness. Not only was the algorithm able to compute the similarity of the drug-pair based on the hypothesis that a drug-pair is similar if the cosine similarity ratio between the drug-pair is high, but such information can also be adopted as a decision support tool for the health professional in drug prescription.

4.6.1 Model Performance

Table 5 shows the results obtained from individual models by running the experiment with the two sets of drug-pairs.

The word embedding model had a higher F score in predicting positive drug-pairs, hence leading to the higher recall rate of 0.85 compared to 0.61 for the adverse network model. In contrast, it had the lower precision rate (0.67 against 0.94), which measured the fraction of positive records that were accurately predicted. This was due to an increase in the false predictions (number of false positives). As the true positives increase, the number of positive pairs that were not correctly predicted (false negatives) decreases, resulting in an increase in the recall rate. When common paths for those drugs in adverse interaction with the original set of interactive drugs are included, the F score dropped drastically compared to the case when radius $= 1$ where only the set of interactive drugs with the vertex was considered. As expected, the performance deteriorated when additional attributes of adverse interactions, such as minor and moderate interactions, were introduced. However, due to fewer possible

Table 5 F score distribution

Threshold	Adverse network				Word Embedding		
	r = 1 (Major)	r = 1 (All)	r = 2 (Major)	r = 2 (All)	w = 2 L = 16	w = 4 L = 16	w = 4 L = 8
0.1	0.55	0.61	0.52	0.70	0.67	0.63	0.65
0.2	0.51	0.55	0.45	0.59	0.67	0.64	0.65
0.3	0.57	0.42	0.41	0.49	0.67	0.71	0.66
0.4	0.68	0.44	0.40	0.44	0.66	0.74	0.68
0.5	0.74	0.43	0.47	0.36	0.64	0.75	0.71
0.6	0.74	0.43	—	0.35	0.61	0.71	0.70
0.7	0.71	0.41	–	0.38	0.51	0.62	0.74
0.8	0.68	0.39	–	0.34	0.31	0.36	0.60
0.9	0.67	0.39	–	–	–	–	–

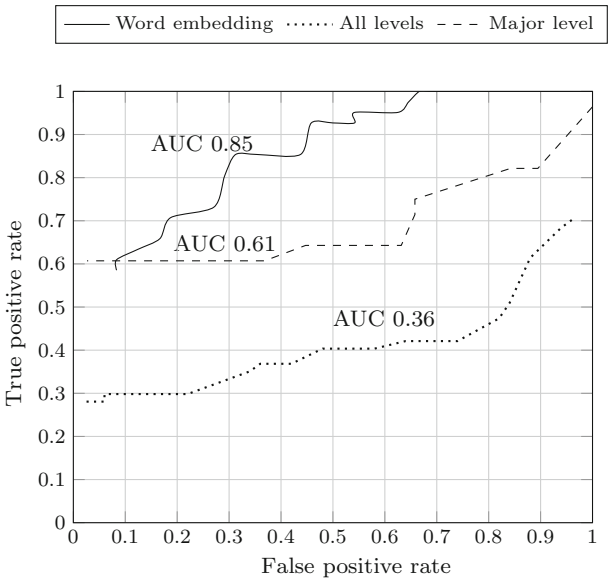

Fig. 6 Comparing AUC for different models

paths when only major interactions are considered, the threshold occurs sooner, where beyond that, there were no true positives obtained in the experiment, which explains the unavailability of F score when the cut-off was over 0.6. With the word embedding model, F score was at a maximum at a layer size of 16. Performance deteriorated when the layer size was decreased since important information from the drug corpus was lost. Window size also affected the F score. Since the number of words before and after the target word was decreased, the quality of the training model is adversely affected, hence the drop in performance with a smaller window size.

Since the precision does not factor in the correct negative predictions within the drug-pairs, (the true negatives, TN), we attempt to assess this performance by plotting the true positive

Table 6 Effect of proximity and nodes properties on performance of the adverse network model

Property	Promixity	Recall	Precision	Accuracy	F score
Major only	1	0.61	0.94	0.82	0.74
	2	0.34	0.75	0.60	0.47
Combined	1	0.30	0.74	0.57	0.43
	2	0.34	0.38	0.37	0.36

Table 7 Influence on performance by training parameters

Window size	Layer size	Recall	Precision	Accuracy	F score
2	8	1.00	0.49	0.52	0.66
2	16	0.98	0.49	0.53	0.66
4	8	0.98	0.56	0.63	0.71
4	16	0.85	0.67	0.74	0.75

rate *tpr* against the false positive rate *fpr* to obtain the receiver operating characteristic (ROC) curve ([7]). With this plot, the area under curve (AUC) can be used to further determine the performance of the model in a more comprehensive manner. A higher AUC indicates a better performance ([6]). The AUC for the word embedding model is 0.85 compared with that for the network model which is 0.61 (Fig. 6). When minor and moderate interactions were also included in considering the number of common paths within the drug-pair, it was noted from the ROC that the AUC was less than 0.5. This is due to the noise introduced into the experiment with the additional paths, which does not aid performance.

4.6.2 Sensitivity Study

Experimental parameters were varied to find the combination that yielded the best performance. These parameters included the proximity distance from the root node and the property of the relationship between the nodes in the adverse network model. Word size and layer size were also varied in the word embedding model. As shown in Table 6, the adverse network model performed best by only considering the major relationship between nodes in the immediate neighbourhood of each drug in the drug-pair. This was the setting used in comparing the performance of the two models in Sect. 4.6.1.

Table 7 shows the performance of the word embedding model with varying window sizes and layer sizes. Changing the window size affected the performance significantly.

Since a smaller number of words before and after the target word was used during training, it is expected that the probability of a word match with the drug-pair during the experiment would lower, hence the drop in performance. Changing the layer size had minimal impact on the performance. The model performed best at a window size of 4 and a layer size of 16.

4.6.3 Drug Prescription Scenario Using Our Model

The framework described in this paper can be easily used in a typical clinical environment to assist the health practitioner in drug prescription at point-of-care. This will ensure the drug

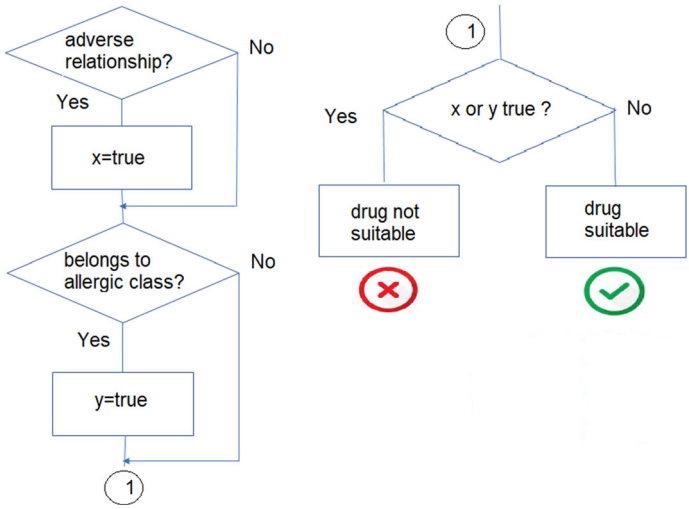

Fig. 7 Using the model in a clinical settings

is not in adverse relationship with what the patient is taking, as well as dissimilar to the drugs that the patient is allergic to.

As illustrated in Fig. 7, such a clinical decision support system consists of two tests. The first test is to ensure the drug to be prescribed is not in adverse relationship with the drug the patient is currently taking. Based on the relationship among drug pairs in the drug taxonomy, the system will search for any adverse relationship between the drug to be prescribed and each drug that the patient is currently taking.

The second test is to ensure the drug to be prescribed does not belong to the same class as the drug that the patient is allergic to. If the drug to be prescribed is either in adverse relationship with the drug that the patient is currently taking or belongs to the same class as the drug that the patient is allergic to, the system will advise the user through the presentation layer that the drug is not suitable and will thus recommend an alternative drug. On receiving the suggestion of the alternative drug, it is then for the user to decide whether this is an appropriate drug to prescribe after further consideration of the duration and dosage of the patient's current drugs.

5 Conclusions

In this paper, a three-tiered conceptual framework is described which enables the similarity ratio of a drug-pair to be computed using feature vectors constructed from bio-medical text. This similarity ratio can then be used to decide if a drug-pair is suitable for prescription. Two different approaches were used to obtain the feature vectors, the word embedding model and the adverse network model. Experimental results showed better performance with the word embedding approach. We have also shown how our framework can be adopted at point-of-care within a CDSS for safe drug prescription by considering the patient's personal medical profile. Other extensions to the models can also be explored. One of these tools is Glove [21], an approach combining the local word embedding method of Word2Vec with the global statistics of matrix factorisation techniques. As Semantic Web allows the data to be

represented in a different format [8], it will be exciting to see the performance of the models by leveraging such technology with the data from DrugBank. Besides DrugBank, it will be interesting to conduct the experiment with alternative data repositories such as PubMed (http://www.pubmed.gov/) and compare the results to evaluate if it is more efficient. In order to optimize the performance of the experiment, it is proposed that individual models described in this paper be amalgamated to form an ensemble model.

With the breakthrough in using similarity ratios within a personalised CDSS, our work will provide further motivation in developing other approaches for determining the similarity ratio of a drug-pair and to extend the use of such a system to pharmaceutical domains.

References

1. Ayvaz S, Horn J, Hassanzadeh O, Zhu Q, Stan J, Tatonetti NP, Vilar S, Brochhausen M, Samwald M, Rastegar-Mojarad M, Dumontier M, Boyce RD (2015) Toward a complete dataset of drug–drug interaction information from publicly available sources. Biomed Inform 55:206–217
2. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
3. Bokharaeian B, Diaz A, Chitsaz H (2016) Enhancing extraction of drug–drug interaction from literature using neutral candidates, negation, and clause dependency. PLoS ONE 11(10):1–20. https://doi.org/10.1371/journal.pone.0163480
4. Bui Q, Sloot P, vanMulligen E, Kors J (2014) A novel feature-based approach to extract drug–drug interactions from biomedical text. BioInformatics 30(23):3365–3371
5. Casillas A, Prez A, Oronoz M, Gojenola K, Santiso S (2016) Learning to extract adverse drug reaction events from electronic health records in spanish. Exp Syst Appl 61:235–245
6. Chen L, Fang B, Shang Z, Tang Y (2018) Tackling class overlap and imbalance problems in software defect prediction. Software Quality Journal 26(1):97–125
7. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874
8. Gheisari M, Movassagh A, Qin Y, Yong J, Tao X, Zhang J, Shen H (2016) Nsssd: a new semantic hierarchical storage for sensor data. In: Proceedings of the 2016 IEEE 20th international conference on computer supported cooperative work in design, CSCWD 2016, pp. 174–179. Institute of Electrical and Electronics Engineers Inc
9. Goh EZ, Beech N, Johnson NR (2020) Dental anxiety in adult patients treated by dental students: a systematic review. J Dent Educ. https://doi.org/10.1002/jdd.12173
10. Goh WP, Tao X, Zhang J, Yong J (2016) Decision support systems for adoption in dental clinics: a survey. Knowl Based Syst 104:195–206
11. Goh WP, Tao X, Zhang J, Yong J, Qin Y, Goh EZ, Hu A (2018) Exploring the use of a network model in drug prescription support for dental clinics. In: The 5th international conference on behavioral, economic, and socio-cultural computing, 12–14 Nov 2018, Kaohsiung, Taiwan
12. Jeh G, Widom J (2002) Simrank: A measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '02, pp 538–543. ACM, New York, NY
13. Khalilfa M (2014) Clinical decision support: strategies for success. Proc Comput Sci 37:422–427
14. Lafta R, Zhang J, Tao X, Li Y, Tseng VS, Luo Y, Chen F (2016) An intelligent recommender system based on predictive analysis in telehealthcare environment. Web Intell 14:325–336
15. Lai S, Liu K, He S, Zhao J (2018) How to generate a good word embedding? IEEE Intell Syst 1–1
16. Levy O, Goldberg Y, Dagan I (2015) Improving distributional similarity with lessons learned from word embeddings. Trans Assoc Comput Linguist 3:211–225
17. Liu S, Tang B, Chen Q, Wang X (2016) Drug–drug interaction extraction via convolutional neural networks. Comput Math Methods Med 2016:1–8
18. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Bengio Y, LeCun Y (eds) 1st International conference on learning representations, 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings
19. Naili M, Chaibi AH, Ghezala HHB (2017) Comparative study of word embedding methods in topic segmentation. In: Procedia computer science **112**, 340–349. Knowledge-based and intelligent information & engineering systems: proceedings of the 21st international conference, KES-2017 6-8 September 2017, Marseille, France

20. Palma G, Vidal ME, Raschid L (2014) Drug-target interaction prediction using semantic similarity and edge partitioning. In: Mika P, Tudorache T, Bernstein A, Welty C, Knoblock C, Vrandečić D, Groth P, Noy N, Janowicz K, Goble C (eds) The Semantic Web—ISWC 2014. Springer, Cham, pp 131–146
21. Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar. https://doi.org/10.3115/v1/D14-1162
22. Pozzi A, Gallelli L (2011) Pain management for dentists: the role of ibuprofen. Annali di Stomatologia 2(3–4 Suppl):3–24
23. Shastri S, Mansotra V (2019) Kdd-based decision making: a conceptual framework model for maternal health and child immunization databases. In: Advances in computer communication and computational sciences, pp 243–253
24. Shi C, Li Y, Zhang J, Sun Y, Yu P (2017) A survey of heterogeneous information network analysis. IEEE Trans Knowl Data Eng 29:17–37
25. Shi C, Zhang Z, Luo P, Yu PS, Yue Y, Wu B (2015) Semantic path based personalized recommendation on weighted heterogeneous information networks. In: Proceedings of the 24th ACM international on conference on information and knowledge management, CIKM '15, pp 453–462. ACM, New York
26. Suda K, Henschel H, Patel U (2018) Use of antibiotic prophylaxis for tooth extractions, dental implants, and periodontal surgical procedures. Open Forum Infect Dis 5(1):250–254
27. Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B (2014) Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers), pp. 1555–1565. Association for Computational Linguistics, Baltimore, Maryland
28. Tari L, Anwar S, Liang S, Cai J, Baral C (2010) Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics 26(18):547–553
29. Wang C, Song Y, Li H, Sun Y, Zhang M, Han J (2017) Distant meta-path similarities for text-based heterogeneous information networks. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 1629–1638. ACM, New York
30. Wang Y, Liu S, Rastegar-Mojarad M, Wang L, Shen F, Liu F, Liu H (2017) Dependency and AMR embeddings for drug–drug interaction extraction from biomedical literature. In: Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics, ACM-BCB '17, pp. 36–43. ACM, New York. https://doi.org/10.1145/3107411.3107426
31. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
32. Yan S, Jiang X, Chen Y (2013) Text mining driven drug–drug interaction detection. In: 2013 IEEE international conference on bioinformatics and biomedicine, pp 349–354
33. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) In: Advances in neural information processing systems 28, pp 649–657. Curran Associates, Inc
34. Zhang Y, Jatowt A, Tanaka K (2016) Towards understanding word embeddings: automatically explaining similarity of terms. In: 2016 IEEE international conference on big data (big data), pp 823–832
35. Zhao Z, Yang Z, Ling L, Lin H, Jian W (2016) Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics 32(22):3444–3453
36. Zhu Y, Yan E, Wang F (2017) Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. BMC Med Inform Decis Mak 17(1):95