Journal of Computer Science

Original Research Paper

Adverse Drug Reaction Detection Using Latent Semantic Analysis

Ahmed Adil Nafea, Nazlia Omar and Mohammed M. AL-Ani

Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

Article history
Received: 23-05-2021
Revised: 27-09-2021
Accepted: 04-10-2021

Corresponding Author:
Ahmed Adil Nafea
Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia
Email: ahmed.adil.nafea@gmail.com

Abstract: Detecting Adverse Drug Reactions (ADRs) is one of the important information for determining the view of the patient on one drug. Most studies have investigated the extraction of ADRs from social networks, in which users share their opinion on a particular medication. Some studies have used trigger terms to detect ADRs. Such studies showed remarkable performance in terms of extracting ADR. However, these terms only would not be sufficient since it needs to be extended periodically when new side effects or new medical-related entities are being discovered. In addition, the feature space with trigger terms would lack latent semantic. This study aims to propose a semantic method based on Latent Semantic Analysis (LSA) for improving the detection of ADR. A benchmark dataset has been used in the experiments along with several pre-processing operations that have been applied including stop word removal, tokenization and stemming with three classifiers that were trained on the proposed LSA, namely Support Vector Machine (SVM), Naïve Bayes (NB) and Linear Regression (LR). In addition, two representations of documents were used, namely Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). Results showed that the proposed LSA outperformed the baseline extended trigger terms by achieving 82% of F-measure for the dataset. Such superiority highlights the use of LSA where the semantic correspondences could be identified correctly rather than using a predefined list of trigger terms.

Keywords: Adverse Drug Reaction, Latent Semantic Analysis, Naïve Bayes, Support Vector Machine, Linear Regression

Introduction

The rise of social networks has contributed toward expanding the textual information dramatically in the last years. Regular users nowadays would have the ability to freely express their minds toward plenty of subjects (Kiritchenko *et al*., 2018; Yousef *et al*., 2019). One of these subjects is the product review where a user can evaluate a specific product and describing its advantages and disadvantages based on his/her experience with the product (Liu *et al*., 2017). ADR detection has been depicted in the literature where numerous studies have crawled data from social networks such as Twitter or from drug websites. In such data collection, the comments or reviews by regular users have been addressed in order to extract the ADR mentions. For example, a review of 'after I took this medicine, I felt dizzy' contains an ADR of 'dizzy' where the user in this review is describing a side-effect from taking a particular medicine.

Several studies proposed different techniques for ADR extraction. Most of the studies have utilized machine learning technique classifiers such as SVM and NB. For the feature space, most of the studies have used the trigger terms (Ebrahimi *et al*., 2016; Kiritchenko *et al*., 2018; Pain *et al*., 2016; Plachouras *et al*., 2016; Yousef *et al*., 2019). Yet, using trigger terms only would not be sufficient since it needs to be extended periodically when new side effects or new medical-related entities are being discovered. In addition, the feature space with trigger terms would lack latent semantic. For example, 'I took this medicine' and 'I consume pills' both sentences have trigger terms of 'took' and 'consume'. Examining the two words in the feature space would be ineffective since they have the same meaning. Therefore, it is necessary to address a semantic technique for improving detection accuracy. In fact, examining the semantic aspect would require the use of an external knowledge source. With the demand for building a specific knowledge source for the adverse drug reaction, the challenge becomes harder. Therefore, it is essential to address a technique that can utilize the semantic aspect without the use of external knowledge sources. Such a technique could be the Latent

Semantic Analysis (LSA) where the semantic correspondences can be determined statistically. This study has proposed the LSA approach for improving the process of identifying whether the sentence has ADR or not in social reviews. Such identification would facilitate the discovery of new side effects of new medicines from regular people through social media and its impact on people's health.

The aim of this study is to propose a semantic method based on LSA for improving the detection of ADR. A benchmark dataset has been used in the experiments along with several pre-processing operations that have been applied including stop word removal, tokenization and stemming with three classifiers that were trained on the proposed LSA, namely SVM, NB and LR. In addition, two representations of documents were used namely TF and TF-IDF. When integrated into a medical opinion mining system, the result of this study can help not only patients assess the drug before taking it, but also doctors and drug producer organizations to consider user feedback in their decision-making process. This algorithm is also applicable to pharmacovigilance systems. In this study we construct the paper as follows: In the section II, we discuss the related works. Following that, we present our proposed method in section III. In section IV, we explain the experimental results and discussion. We complete our findings in section V with decisive outcomes and rational future recommendations

Related Work

The literature has shown great interest in the task of ADR detection. The benchmark dataset of medical reviews was first presented by Yates and Goharian (2013). These authors also utilized trigger terms with the rule-based technique to identify the studies with ADR. This study proposed the extraction of ADR automatically from user feedback on different social media platforms to classify adverse reactions not reported by the United States Food and Drug Administration (FDA). This proposal utilised different lexicons, identification patterns and created a range of synonyms, including variations in medical terminology and identification trends. They identify "expected" and "unexpected" ADRs. The context language (drug) was used to determine the frequency of unexpected, detected ADR.

Pain *et al*. (2016) presented an ADR detection technique using SVM to the classifier. The proposed method utilized a set of keywords and hashtags trigger terms that were frequently occurring with ADR. The authors used a medical review of collected data from Twitter to provide automatic drug-effect detection. The proposed features can identify numerous types of drug-effect entities. Their research described developing Post-Marketing Surveillance (PMS) methods specifically in particular for messy types of text found on Twitter.

Ebrahimi *et al*. (2016) employed a set of medical concepts with specifically named entities as trigger terms to determine the side effects of drugs from medical reviews. POS tagging was utilized to identify the syntactic tag of terms. Two classifiers, namely, a rule-based classification method and SVM, were adopted to detect the side effects of drugs. This research developed a method to identify side effects in medication reports as a subtask to identify implicit perceptions in medical literature and distinguish side effects and disease symptoms.

Plachouras *et al*. (2016) applied a set of trigger terms or gazetteer features, along with an N-gram representation, to extract adverse drug events from Twitter reviews. The research presented a system for large-scale pharmacovigilance support. The authors tackled the question of adverse event extraction from tweets via training and testing a supervised binary classifier. SVM classification method was implemented by the authors to accommodate the final extraction by using words and keywords, surface characteristics, a list of gazetteers, POS tags and sentiment analysis.

A group of researchers from NRC-Canada Kiritchenko *et al*. (2018) at the AMIA-2017 Workshop on Social Media Mining for Health Applications (SMM4H), engaged in two joint activities. The first activity, Task 1, was about classifying tweets with reference to ADR, while Task 2 focused on classifying tweets describing personal intake of medications. With regard to both tasks, vector machine classifiers were trained using a variety of surface-specific features, feelings and domain-specific features through the presentation of an SVM technique for ADR extraction. The authors filtered the trigger terms to use a domain-specific one for improving the accuracy of detection. Experiments were conducted using Twitter medical reviews.

Emadzadeh *et al*. (2017) has used latent semantic analysis with a hybrid semantic analysis in order to combine the Unified Medical Language System (UMLS) to improve the performance in terms of extracting ADR. In regard to their corresponding standardized identifiers, this study proposed a modular NLP pipeline for mapping (normalizing) colloquial mention of ADRs. For evaluation, they use a publicly available, annotated corpus of 2008 tweets (Nikfarjam *et al*., 2015).

The study of Yousef *et al*. (2019) tackled the extraction of ADR from social networks where users express their views on a specific medication. Obtaining entities mainly depends on specific terms that may occur before or after ADR, called trigger terms. However, those terms should be constantly extended, modified and updated. The aim of this study was to propose an extension of the trigger terms based on the multiple N-gram representations. Two document representations including the TF-IDF and TF were used. The experiments were conducted using secondary data from drug websites.

Most techniques utilize annotated data of medical review in order to train a classification model. Within the training, there are several features that can be used to indicate the occurrence of ADR. One of these features is the trigger terms which are the keywords that are frequently accompanied by

ADRs. Researchers have extensively used this type of feature with different classification methods. Consequently, the key limitation behind their studies lies in the dependency of using trigger terms where the semantic aspect could be discarded. The novelty of this study is represented by using LSA instead of trigger terms which have been examined by the literature (Ebrahimi *et al*., 2016; Kiritchenko *et al*., 2018; Pain *et al*., 2016; Plachouras *et al*., 2016; Yousef *et al*., 2019). LSA is a technique that has been used for identifying the semantics of terms statistically (Al-Sabahi *et al*., 2018). Furthermore, unlike trigger terms which have been intended to filter the Bag of Word search space in order to maintain significant terms, LSA will identify the semantic correspondences without losing any important information. LSA, therefore, has the ability to configure the meaning of terms based on the similarity of contexts.

One of the state-of-the-art semantic approaches like Cocos *et al.* (2017) have used a deep learning approach of RNN to extract ADR based on the embedding of words. On the other hand, there is another study by Liu and Lee (2018) which used CNN for generating the word embedding to detect ADR. This study applied CNN with a number of classifiers, for example when CNN was applied with a series of classifiers such as Radial Basis Function Neural Network (RBFNN), Logistic Regression (LR), Multilayer Perceptron (MLP) and SVM. However, the authors used different Twitter dataset. The key limitation behind their studies lies in the dependency of using word embedding focused on the term sequences and it needs to pretrain the model where the word embedding focused on the term sequences, but LSA does not focus on this. LSA utilizes the statistical information and implicitly identifies the most important trigger terms. This enhances the identification of the semantic connection and relationship between the terms. This finding implies the effectiveness of proposing LSA of extracting ADR.

Proposed Methods

The methodology of this study consists of five phases as shown in Fig. 1. The first phase is the preparation of annotated drug reviews where the dataset used is from a benchmark dataset by Yates and Goharian (2013) in which Yousef *et al.* (2019) modified some of some structure by adding more meaningful columns of the data. The second phase will contain pre-processing tasks such as tokenization, stop word removal and stemming. The third phase aims to represent the terms in a vector space representation using both TF and TF-IDF. The fourth phase contains the semantic analysis using the proposed LSA. The fifth phase will address the classification where three classifiers will be used including SVM, NB and LR. Each phase is discussed in further detail in the next subsections.

Dataset

The dataset used is from a benchmark dataset by Yates and Goharian (2013) in which Yousef *et al*. (2019) modified some of some structure by adding more meaningful columns of the data. The original data set contains 3 columns (Doc, ADR and review) and after being modified it consisted of five columns (Doc, Sen, Class, Review and ADR). The dataset used in this study contains 2500 reviews (with 246 labeled documents). Each document contains one or more sentences. The documents contain 944 sentences in total. Those sentences are collected from Twitter platform. The total number of ADR are 982 for all documents. These documents are written in the English language. The review dataset is collected from Drug Review Sites on social media, namely, drugratingz. com, askapatient. com and drugs.com.

Table 1 shows the dataset details and Table 2 shows a sample example of dataset

Preprocessing

In this stage, the process of splitting the text when running on a set of pre-processing algorithms to prepare it for the next stages. The above tasks can be described as follows.

Stop word removal: This activity is aimed at eliminating a language's common words that don't hold any important details of their own. At the pre-processing point, these terms are often omitted to reduce the number of less informative features known as noise data (Kaur and Buttar, 2018; Oliinyk *et al*., 2020). Figure 2 shows an example of stripping of the stop-words.

Tokenization: Is a process that attempts to transform the text into a sequence of sentences and then convert those sentences into sequences of tokens (i.e., words) (Chary *et al*., 2019). Figure 3 shows the tokenization process.

Stemming: The final stemming preprocessing step will be applied. This mission aims at restoring the origin of words by removing the various suffixes. In this study, Porter's Stemmer algorithm (Porter, 1980) was used for this manner It is based on the idea that suffixes in English (Patel and Passi, 2020). Figure 4 shows an example of a function with stemming words.

Term Representation

In this stage, the data will be represented the number of occurrences of the word in the documents by the TF or TF-IDF.

Term Frequency (TF): - In this process, the number of occurrences of the word in the document is represented at the TF. The formula used to solve the problem concerning frequency is:

$$W_d(t) = TD(t,d) \quad (1)$$

where, $TD(t, d)$ is the word T frequency in document d.

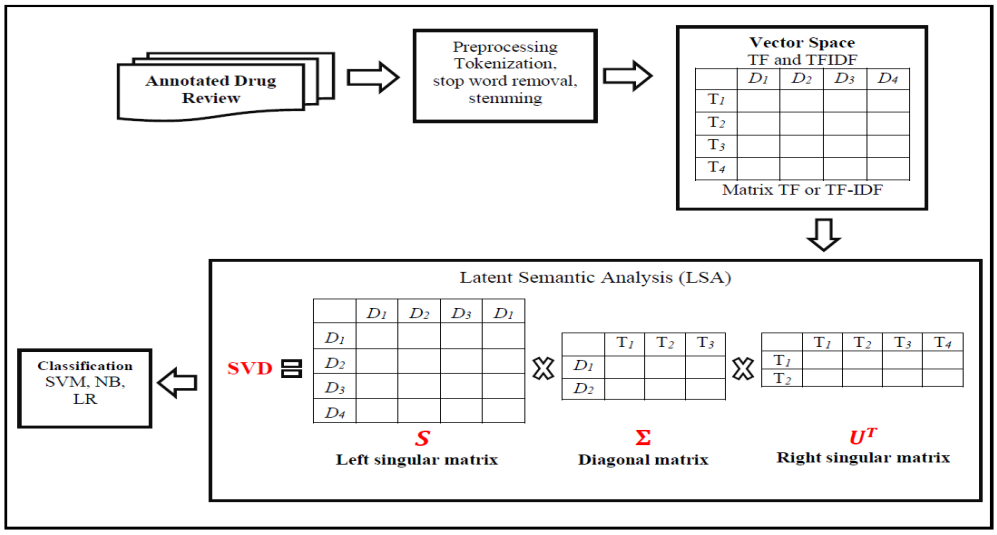

Fig. 1: Proposed LSA methods

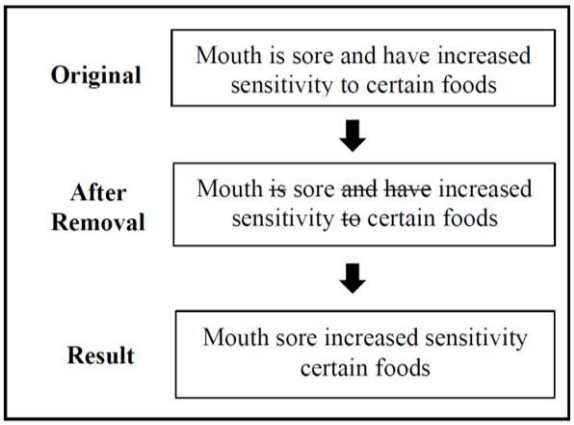

Fig. 2: Example of removing stop words

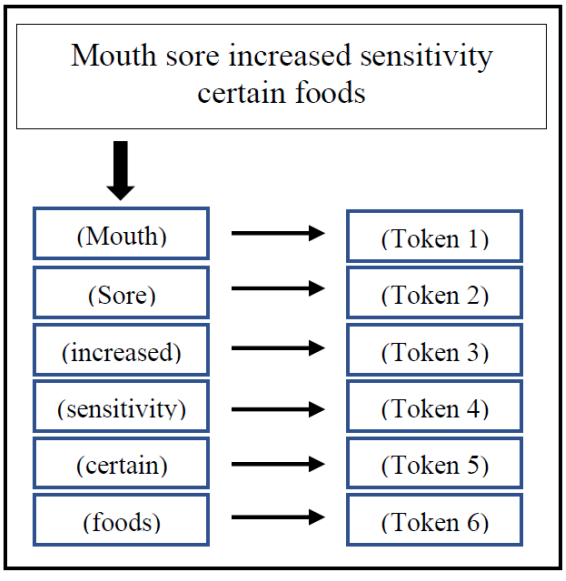

Fig. 3: Example of tokenization

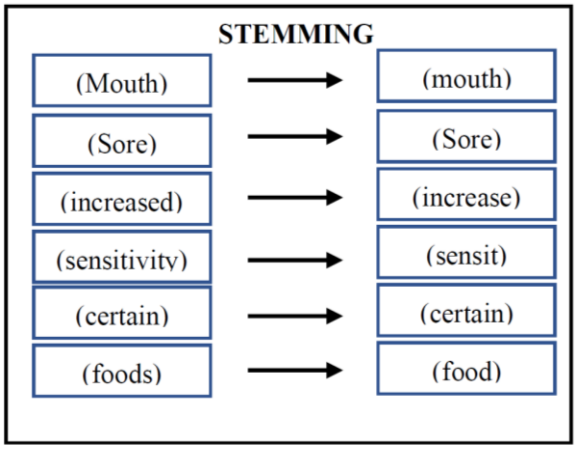

Fig. 4: Example of the stemming process

Inverse Document Frequency (*IDF*):- *IDF* seeks to have high weight for unusual conditions and low typical conditions weights. The formula reads as:

$$IDF_t = \log n\left(\frac{N}{N_t}\right) \quad (2)$$

where, N_t is the number of documents that contain the word and where N is the number of English documents.

Term Frequency with Inverse Document Frequency TF-IDF: - This method is a combination of two preceding TF and IDF methods. The formula regarding weighting as follows (Chen *et al*., 2016; Mohammed and Omar, 2020):

$$W_t = TF(t,d).IDF_t \quad (3)$$

where, *TF* (*t*, *d*) refers to the Term Frequency *t* in document d and IDF_t (refers to the inverse document frequency of term *t*.

Table 1: Dataset details

Attribute	Total
Number of total reviews	2500 (labeled 246)
Number of sentences	944
Number of ADR	982

Table 2: Sample of the dataset

Doc	Sen	Class	Review	ADR
1	1	1	My joint pain is very severe.	['pain']
2	1	0	I was fine in the beginning.	[]
2	2	1	Lower back pain.	['pain']
2	3	0	Swelling of hands.	[]
3	1	1	General Muscle Aches and Fatigue.	['fatigue']
4	1	1	Numbness in toes	['Numbness']
4	2	1	Can't walk, everything aches.	['aches']

Table 3: Example of three documents

	Sample of medical text documents
D1 =	Shoot pain knee feet
D2 =	Infrequ joint pain
D3 =	Experience severe joint pain

Table 4: Calculating the TF

Words	D_1	D_2	D_3
Shoot	1	0	0
Knee	1	0	0
Feet	1	0	0
Infrequ	0	1	0
Experience	0	0	1
Severe	0	0	1
Joint	0	1	1
Pain	1	1	1

Table 5: IDF Calculation

Words	IDF
Shoot	$\log_2\left(\frac{3}{1}\right)=0.477$
Knee	$\log_2\left(\frac{3}{1}\right)=0.477$
Feet	$\log_2\left(\frac{3}{1}\right)=0.477$
Infrequ	$\log_2\left(\frac{3}{1}\right)=0.477$
Experience	$\log_2\left(\frac{3}{1}\right)=0.477$
Severe	$\log_2\left(\frac{3}{1}\right)=0.477$
Joint	$\log_2\left(\frac{3}{2}\right)=0.176$
Pain	$\log_2\left(\frac{3}{1}\right)=0$

Table 6: TF-IDF calculation

Words	D_1	D_2	D_3
Shoot	0.477	0	0
Knee	0.477	0	0
Feet	0.477	0	0
Infrequ	0	0.477	0
Experience	0	0	0.477
Severe	0	0	0.477
Joint	0	0.176	0.176
Pain	0	0	0

For example, consider three statements (i.e., documents) D1, D2 and D3, which have sentences as shown in Table 3.

To calculate the TF, first be determined for every word found in the statements. The singular terms are segmented as in Table 4.

As shown in Table 4 the value 1 is the word present in the phrase corresponding to the sentence given, while 0 is the absence of the word corresponding to the statements given.

Therefore, IDF will be calculated for each word corresponding to the specified three documents, note that N = total number of documents which is 3 and N_t is the number of word appearances in the three documents. IDF for each term can be determined based on Eq. (2), as shown in Table 5.

Finally, by multiplying the TF and IDF, TF-IDF can be obtained. This multiplication is shown in Table 6.

Proposed Latent Semantic Analysis

The Latent Semantic Analysis is a technique commonly used in the processing of NLP to define the similarities between two text classes (Froud *et al*., 2013; Mezher and Omar, 2016). It attempts to analyze the relationships between two sets of documents by constructing a vector space for the meanings of both documents' phrases, expressions and concepts. It can be achieved by vectoring the terms into two rows and columns where the terms are displayed in the rows and the documents in the columns represented. Using the frequency principle of terms theory, LSA can determine the essential relationship by counting the frequency of terms (Islam and Hoque, 2010). Given the high dimensionality of the words in question, a post-processing technique called Singular Value Decomposition (SVD) is applied to minimize the dimensionality of the word matrix. In particular, SVD aims to reduce the number of rows without losing the structure of similarity between columns (Filieri *et al*., 2021; Manning and Schutze, 1999). Basically, LSA implements the matrix using TF or TFIDF by identifying the occurrences of words in respect documents. Hence, the Singular Value Decomposition (SVD) is applied in order to reduce the dimensionality of the word vector. The following equation can be used for calculating SVD:

$$SVD = S\Sigma U^T \quad (4)$$

where, S is the left singular matrix, Σ is a diagonal matrix and U^T is the right singular matrix.

This is conducted through a process known as Singular Value Decomposition (SVD). Then it will be classified by one of the classifications (SVM, NB and LR) that he used in the baseline (Yousef *et al*., 2019).

To illustrate the SVD, let X be an array containing three sentences for D1 and D2 with D3 which are the dataset statements used as shown in Table 3.

This is a simple example of the work of the LSA. The *TF* representation has been stated as in Table 4.

In order to get the SVD, *Y* has to be calculated where *Y* is the union of documents in terms of words $Y = X^T$* X where X^T is the transpose of *X*. In addition, *Z* has to be calculated where *Z* is the union of words in terms of documents $Z = X * X^T$. First, the matrix *X* and its transpose X^T will be represented as follow:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} X^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Since $Y = X\,X^T$, so it can be represented as follow:

$$Y = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Hence, the results of the previous multiplication will be equivalent as follow:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 2 & 3 \end{bmatrix}$$

Similarly, $Z = X^T X$, so it can be calculated as follow:

$$Z = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 4 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 4 \end{bmatrix}$$

Therefore, to compute the SVD, using the Eq. (4) following formula has to be applied:

$$SVD(X) = S\Sigma Y^T$$

where, *S* is the eigenvector of *Y* and *U* is the eigenvector of *Z* and *Σ* is the root square of the eigenvalue of *Z*.

$$Eigenvector\ of\ Y = S = \begin{bmatrix} 0.29511 & 0.000i \\ 0.29511 & 0.000i \\ 0.29511 & 0.000i \\ 0.31639 & 0.000i \\ 0.38848 & 0.000i \\ 0.38848 & 0.000i \\ 0.70488 & 0.000i \\ 1 & 0.000i \end{bmatrix}$$

$$Eigenvector\ of\ Z = U = \begin{bmatrix} 0.75965 & 0.000i \\ 0.81442 & 0.000i \\ 1 & 0.000i \end{bmatrix}$$

$$Transpos\ of\ (U) = U^T = \begin{bmatrix} 0.75965 & 0.81442 & 1 \\ 0.000 & 0.000 & 0.000 \end{bmatrix}$$

$$Eigenvalue\ of\ Z = \begin{bmatrix} 6.3885 \\ 3.0873 \\ 1.4242 \end{bmatrix}$$

$$\sum = \begin{bmatrix} \sqrt{6.3885} & 0 & 0 \\ 0 & \sqrt{3.1873} & 0 \\ 0 & 0 & \sqrt{1.4242} \end{bmatrix} = \begin{bmatrix} 2.52 & 0 & 0 \\ 0 & 1.78 & 0 \\ 0 & 0 & 1.19 \end{bmatrix}$$

$$SVD(X) = S\Sigma U^T = \begin{bmatrix} 0.56493438498 & 0.427811405436 & 0.3511809 \\ 0.56493438498 & 0.427811405436 & 0.3511809 \\ 0.56493438498 & 0.427811405436 & 0.3511809 \\ 0.60567107202 & 0.458660331964 & 0.3765041 \\ 0.74367425664 & 0.563166869248 & 0.4622912 \\ 0.74367425664 & 0.563166869248 & 0.4622912 \\ 1.34936447184 & 1.021841697888 & 0.8388072 \\ 1.914318 & 1.4496676 & 1.19 \end{bmatrix}$$

Here the complex matrix is completed in finding the semantic. LSA first utilizes either TF or TFIDF where all the unique words are grouped in separated attributes. Hence, LSA inputs either CV or TFIDF matrix and outputs the same dimension matrix but with more sophisticated values that adequately indicate the semantic behind every term. This is conducted through a process known as Singular Value Decomposition (SVD). Then it will be classified by one of the classifications (SVM, NB and LR) that is used in the baseline (Yousef *et al*., 2019).

Classification

Machine learning is applied in this step for classifying ADRs. Classification methods like SVM, NB and LR are used to evaluate f-measure efficiency.

Classification methods like SVM, NB and LR are used to evaluate f-measure. The three classifiers are trained on the extracted patterns produced by the proposed LSA. This training aims to build a model that can classify new data in the testing phase. During the training, the model of each classifier learns the cases of the potential occurrence of ADRs. The proposed method is utilized in examining the medical sentiment analysis where it could classify the sentence into 0 (does not have ADR) or 1 (have ADR).

The first method of classification is SVM, which functions by determining an appropriate separator in a 2 dimensional space between data instances. SVM aims at the establishment of the optimal hyperplane with the following decision function (Ebrahimi *et al*., 2016).

$$f(\vec{x}) = \text{sgn}\left(\left(\vec{x} \times \vec{w}\right) + b\right) = \begin{cases} +1: \left(\vec{x} \times \vec{w}\right) + b > 0 \\ -1: \quad Otherwise \end{cases} \tag{5}$$

SVM maps the optimum hyperplane with the optimum margin. Assume a positive and negative data instances partitioned by a hyperplane and the shortest path p+(p-) is lying between the nearest positive and nearest negative instances (Abdullah *et al*., 2009; Hasan and Zakaria, 2016; Moghaddam and Ester, 2011). The margin of this hyperplane, in this case, is given as $p_+ + p_-$.

NB operates by defining the probabilities for the data instances of classes. You can measure the likelihood using the following equation (Elhadad *et al*., 2019; Khalifa and Omar, 2014; Yousef *et al*., 2020).

$$P\left(C_i \mid d\right) = \frac{P\left(C_i\right)P\left(d \mid C_i\right)}{P\left(d\right)} \tag{6}$$

where, given the predictor (*x*, attributes), $P(Ci)$ is the posterior probability of class *Ci*.

LR functions by evaluating the linear class probability equation, which can be seen as follows (Montgomery *et al*., 2015).

$$y = a + bX \tag{7}$$

where, *X* is the dependent variable, the y-intercept is a and b is the line slope.

After implementing the classification (ADR) using the machine learning SVM, NB and LR, it is necessary to validate the results of the categorization performed by the classifier. For the evaluation involving the precision, recall and f-measure, it can be calculated based on the following measures.

Precision

It is a measure of exactness. It is the ration of the predicted positive cases that were correct to the total number of predicted positive cases.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall

Is a measure of completeness. It is the proportion of positive cases that were correctly identified to the total number of positive cases.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where, *TP* is the right classified of *ADR*, *FP* is the wrong classified ADR, FN is the incorrectly rejected classified ADR and TN correctly rejected classified ADR.

F-Measure

It is the harmonic mean of precision and recall. It is an important measure as it gives equal importance to precision and recall:

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

The three classifiers are trained on the extracted patterns produced by the proposed LSA. This training aims to build a model that can classify new data in the testing phase. During the training, the model of each classifier learns the cases of the potential occurrence of ADRs. Table 7 shows the experimental settings.

Experimental Results

Multiple experiments have been conducted to acquire the results for the detection of ADR (i.e., baseline vs. proposed). There are different representations (i.e., TF. TFIDF) and multiple classification methods like SVM, NB and LR which are used to evaluate the performance. The three classifiers are trained on the extracted patterns produced by the proposed LSA as opposed to the baseline research using trigger terms.

As shown in Table 8, the results of f-measure for all classifiers using the proposed LSA via TF with SVM, NB and LR have outperformed the ones by the baseline trigger terms. The performance of SVM based on F-measure has improved from 67% (using trigger terms) to 81% (using LSA). Similarly, the performance of NB based on Fmeasure has improved from 61% (using trigger terms) into 68% (using LSA). Finally, the performance of LR based on F-measure has improved from 67% using trigger terms to 82% (using LSA). Figure 5 displays the f-measure results of the proposed LSA and baseline via TF with SVM, NB and LR classifiers.

As shown in Table 9, the results of f-measure for all classifiers using the proposed LSA via TF-IDF with SVM, NB and LR have outperformed the ones by the baseline trigger terms. The performance of SVM based on Fmeasure has improved from 69% (using trigger terms) to 80% (using LSA). As well as the performance of NB based on F-measure has improved from 61% (using trigger terms) into 72% (using LSA). Finally, the performance of LR based on F-measure has improved from 68% using trigger terms to 80% (using LSA). Figure 6 displays the f-measure results of the proposed LSA and baseline via TF-IDF with SVM, NB and LR classifiers.

On the other hand, Ebrahimi *et al*. (2016) have used SVM to detect ADR using trigger terms and medical concepts as a feature. The performance of SVM based on F-measure has achieved 0.72%. Plachouras *et al*. (2016) have used SVM to extract ADRs using Trigger terms and Gazetteers. The performance of SVM based on Fmeasure has achieved 60.4%. Kiritchenko *et al*. (2018) have used SVM to extract ADRs with domain-specific trigger terms. The performance of SVM based on Fmeasure has achieved 0.68%. The proposed LSA as shown better results compared to other similar researches in terms of detecting ADRs. This finding implies the effectiveness of using LSA in extracting ADRs where the semantic correspondences have been identified correctly rather than using a predefined list of trigger terms.

Such superiority is referred to as the use of LSA where the semantic correspondences have been identified correctly rather than using a predefined list of trigger terms. In a comparison between plain vector space model or the so-called N-gram representation against the feature space generated by LSA, Hutchison *et al*. (2018) have demonstrated better f-measure of classification. This is because LSA can handle synonymy problems within a particular dataset. In addition, LSA can work well on the dataset with diverse topics which exactly would fit the adverse drug reaction datasets where various medical discourses are being tackled.

Apart from the traditional baseline which utilized conventional approaches such as SVM, NB and others, it is necessary to compare the proposed method against state-of-the-art methods that employed deep learning techniques based on word embedding. Liu and Lee (2018) have used CNN for generating the word embedding to detect ADR. they applied deep learning approach of CNN with a lot of methods of machine learning classification They have achieved an f-measure of 58.1%. Lee *et al*. (2017) have used a deep learning approach of CNN to extract ADRs and acquired an f-measure of 64.5%. Cocos *et al*. (2017) have used a deep learning approach of ANN to extract ADRs and acquired an f-measure of 75.5%. The deep learning approaches usually require a large training data for the medical words. The authors used different Twitter data set.

Other studies such as Wang *et al*. (2019) which utilized more sophisticated deep learning approaches have obtained an f-measure of 84.4%. They also used a different dataset. Comparing such results against the proposed method is not possible due to the different dataset. Considering LSA that has been utilized by the proposed method, it is clear that the proposed method is still considered to be less complicated, competitive and simpler in terms of the processes involved and require less computing power.

Table 7: Experimental settings

Experiment	Description
Feature	1. Baseline trigger terms with TF-IDF (Unigram, Bigram, Trigram, and Quadgram)
	2. Baseline trigger terms with Term Frequency (Unigram, Bigram, Trigram, and Quadgram)
	3. Proposed LSA with TF-IDF (Unigram)
	4. Proposed LSA with Term Frequency (Unigram)
Classifiers	1. SVM
	2. NB
	3. LR
Dataset	Benchmark dataset by (Yates and Goharian, 2013) which is then updated by (Yousef *et al*., 2019)
Training and Testing	70% for training and 30% for testing

Table 8: A comparison of results on the proposed approach and baseline based on TF

	SVM	NB	LR
Baseline	67%	61%	67%
Proposed approach	81%	68%	82%

Table 9: A comparison of results on the proposed approach and baseline based on TF-IDF

	SVM	NB	LR
Baseline	69%	61%	68%
Proposed approach	80%	72%	80%

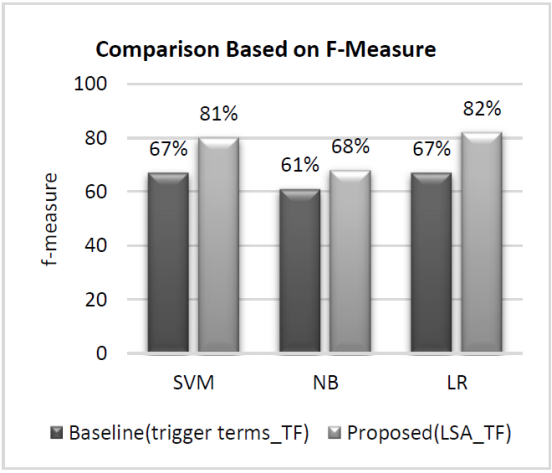

Fig. 5: A comparison of results on the proposed approach and baseline via TF results

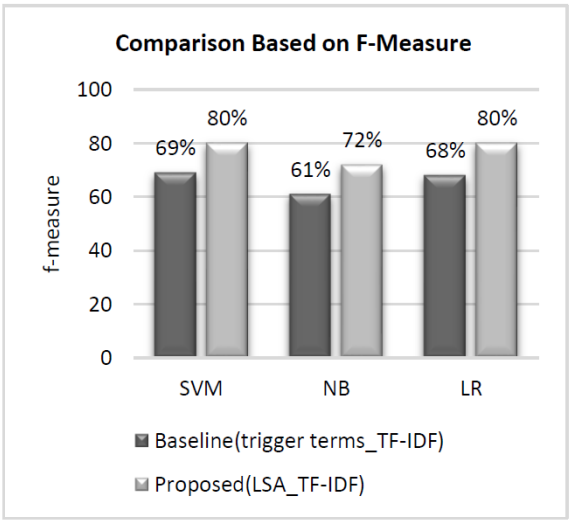

Fig. 6: A comparison of results on the proposed approach and baseline via TF-IDF results

Conclusion

This study proposed an LSA for detecting ADRs. The experiments involved three classifiers, namely, SVM, NB and LR. The proposed LSA achieved higher results than the baseline ones when TF feature and LR classifier was used. This study has proposed an approach for improving the process of identifying whether the sentence has ADR or not through social reviews. Such identification would facilitate the discovery of new side effects of new medicines from regular people through social media and its impact on people's health. To sum up, the proposed LSA has demonstrated competitive performance. This can prove that the use of latent syntactic analysis is playing an essential role in ADR detection. For future work, we plan to experiment with the proposed model upon other ADR datasets, particularly those that have been collected from real-time from social reviews like COVID-19 drug reviews. It is a great challenge since these datasets would contribute to new ADR terms or discover new drug reactions.

Acknowledgement

It is our pleasure to express our sincere gratitude and deepest thanks to our friends in UKM. our valuable discussions, comments and suggestions have greatly improved the content and the presentation of this article. The authors would also like to thank the UKM for funding the research.

Funding Information

This project is funded by UKM under the research code GP-2020-K007009.

Author's Contributions

Ahmed Adil Nafea: Carried out the investigation on techniques of ADR detection and he implemented the proposed work and performed the experiments.

Mohammed M. Al-Ani: Assisted in the realization of the ideas.

Nazlia Omar: Advised the investigation and writing of this manuscript, in which all authors had approved the final version.

Ethics

This article is original and previously unpublished contains material in any journal. The corresponding author acknowledges that the work has been reviewed and approved by all other authors and that there are no ethical concerns.

References

Abdullah, S. N. H. S., Omar, K., Sahran, S., & Khalid, M. (2009, August). License plate recognition based on support vector machine. In 2009 International Conference on Electrical Engineering and Informatics (Vol. 1, pp. 78-82). IEEE. doi.org/10.1109/ICEEI.2009.5254811

Al-Sabahi, K., Zhang, Z., Long, J., & Alwesabi, K. (2018). An enhanced latent semantic analysis approach for arabic document summarization. arXiv preprint arXiv:1807.11618. doi.org/10.1007/s13369-018-3286-z

Chary, M., Parikh, S., Manini, A. F., Boyer, E. W., & Radeos, M. (2019). A review of natural language processing in medical education. Western Journal of Emergency Medicine, 20(1), 78. doi.org/10.5811/westjem.2018.11.39725

Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Systems with Applications, 66, 245-260. doi.org/10.1016/j.eswa.2016.09.009

Cocos, A., Fiks, A. G., & Masino, A. J. (2017). Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. Journal of the American Medical Informatics Association, 24(4), 813-821. doi.org/10.1093/jamia/ocw180

Ebrahimi, M., Yazdavar, A. H., Salim, N., & Eltyeb, S. (2016). Recognition of side effects as implicit-opinion words in drug reviews. Online Information Review. doi.org/10.1108/OIR-06-2015-0208

Elhadad, M. K., Li, K. F., & Gebali, F. (2019, March). Sentiment analysis of Arabic and English tweets. In Workshops of the International Conference on Advanced Information Networking and Applications (pp. 334-348). Springer, Cham. doi.org/10.1007/978-3-030-15035-8_32

Emadzadeh, E., Sarker, A., Nikfarjam, A., & Gonzalez, G. (2017). Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology. In AMIA Annual Symposium Proceedings (Vol. 2017, p. 679). American Medical Informatics Association. https://www.ncbi.nlm.nih.gov/pmc/articles/pmc5977584/

Filieri, R., Galati, F., & Raguseo, E. (2021). The impact of service attributes and category on eWOM helpfulness: An investigation of extremely negative and positive ratings using latent semantic analytics and regression analysis. Computers in Human Behavior, 114, 106527. doi.org/10.1016/j.chb.2020.106527

Froud, H., Lachkar, A., & Ouatik, S. A. (2013). Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering. arXiv preprint arXiv:1302.1612. doi.org/10.5121/ijdkp.2013.3107

Hasan, A. M., & Zakaria, L. Q. (2016). Question classification using support vector machine and pattern matching. Journal of Theoretical and Applied Information Technology, 87(2), 259.

Hutchison, P. D., Daigle, R. J., & George, B. (2018). Application of latent semantic analysis in AIS academic research. International Journal of Accounting Information Systems, 31, 83-96. doi.org/10.1016/j.accinf.2018.09.003

Islam, M. M., & Hoque, A. L. (2010, December). Automated essay scoring using generalized latent semantic analysis. In 2010 13th International Conference on Computer and Information Technology (ICCIT) (pp. 358-363). IEEE. IEEE. doi.org/10.1109/ICCITECHN.2010.5723884

Kaur, J., & Buttar, P. K. (2018). A systematic review on stopword removal algorithms. International Journal on Future Revolution in Computer Science & Communication Engineering, 4(4), 207-210. http://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1499

Khalifa, K., & Omar, N. (2014). A hybrid method using lexicon-based approach and Naive Bayes classifier for Arabic opinion question answering. J. Comput. Sci., 10(10), 1961-1968. doi.org/10.3844/jcssp.2014.1961.1968

Kiritchenko, S., Mohammad, S. M., Morin, J., & de Bruijn, B. (2018). NRC-Canada at SMM4H shared task: classifying Tweets mentioning adverse drug reactions and medication intake. arXiv preprint arXiv:1805.04558. https://arxiv.org/abs/1805.04558

Lee, K., Qadir, A., Hasan, S. A., Datla, V., Prakash, A., Liu, J., & Farri, O. (2017, April). Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In Proceedings of the 26th international conference on world wide web (pp. 705-714). doi.org/10.1145/3038912.3052671

Liu, S., & Lee, I. (2018, October). Sentiment classification with medical word embeddings and sequence representation for drug reviews. In International Conference on Health Information Science (pp. 75-86). Springer, Cham. doi.org/10.1007/978-3-030-01078-2_7

Liu, Y., Bi, J. W., & Fan, Z. P. (2017). Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. Information Fusion, 36, 149-161. doi.org/10.1016/j.inffus.2016.11.012

Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT press.

Mezher, R., & Omar, N. (2016). A hybrid method of syntactic feature and latent semantic analysis for automatic arabic essay scoring. Journal of Applied Sciences, 16(5), 209-215. doi.org/10.3923/jas.2016.209.215

Moghaddam, S., & Ester, M. (2011, December). AQA: aspect-based opinion question answering. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 89-96). IEEE. doi.org/10.1109/ICDMW.2011.34.

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. PloS one, 15(3), e0230442. doi.org/10.1371/journal.pone.0230442

Montgomery, D. C., Jennings, C. L., & Kulahic, M. (2015). Introduction to time series analysis and forecasting. John Wiley & Sons.

Nikfarjam, A., Sarker, A., O'connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association, 22(3), 671-681. doi.org/10.1093/jamia/ocu041

Oliinyk, V. A., Vysotska, V., Burov, Y., Mykich, K., & Fernandes, V. B. (2020). Propaganda Detection in Text Data Based on NLP and Machine Learning. In MoMLeT+ DS (pp. 132-144). http://ceur-ws.org/Vol-2631/paper10.pdf

Pain, J., Levacher, J., Quinquenel, A., & Belz, A. (2016, December). Analysis of Twitter data for postmarketing surveillance in pharmacovigilance. In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT) (pp. 94-101). https://www.aclweb.org/anthology/W16-3914

Patel, R., & Passi, K. (2020). Sentiment analysis on Twitter data of world cup soccer tournament using machine learning. IoT, 1(2), 218-239. doi.org/10.3390/iot1020014

Plachouras, V., Leidner, J. L., & Garrow, A. G. (2016, July). Quantifying self-reported adverse drug events on Twitter: signal and topic analysis. In Proceedings of the 7th 2016 International Conference on Social Media & Society (pp. 1-10). doi.org/10.1145/2930971.2930977

Porter, M. F. (1980). An algorithm for suffix stripping. Program. doi.org/10.1108/eb046814

Wang, C. S., Lin, P. J., Cheng, C. L., Tai, S. H., Yang, Y. H. K., & Chiang, J. H. (2019). Detecting potential adverse drug reactions using a deep neural network model. Journal of medical Internet research, 21(2), e11016. doi.org/10.2196/11016

Yates, A., & Goharian, N. (2013, March). ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In European Conference on Information Retrieval (pp. 816-819). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-36973-5_92

Yousef, R. N. M., Tiun, S., Omar, N., & Alshari, E. M. (2020). Lexicon replacement method using word embedding technique for extracting adverse drug reaction. International Journal of Technology Management and Information System, 2(1), 113-122.

Yousef, R.N.M., Tiun, S., & Omar, N. (2019). Extended trigger terms for extracting Adverse Drug Reactions in social media texts. J. Comput. Sci. 15(6), 873-79. doi.org/10.3844/jcssp.2019.873.879