Original Paper

Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure

Ahmad P Tafti[1], PhD; Jonathan Badger[1], PharmD; Eric LaRose[1], BS; Ehsan Shirzadi[2], MSc; Andrea Mahnke[1], MSc; John Mayer[1], PhD; Zhan Ye[1], PhD; David Page[3], PhD; Peggy Peissig[1], MBA, PhD

[1]Biomedical Informatics Research Center, Marshfield Clinic Research Institute, Marshfield, WI, United States

[2]Institute of Electrical and Electronics Engineers, Dublin, Ireland

[3]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, United States

Corresponding Author:
Ahmad P Tafti, PhD
Biomedical Informatics Research Center
Marshfield Clinic Research Institute
1000 N Oak Ave
Marshfield, WI, 54449
United States
Phone: 1 7153893514
Email: pahlavantafti.ahmad@marshfieldresearch.org

Abstract

Background: The study of adverse drug events (ADEs) is a tenured topic in medical literature. In recent years, increasing numbers of scientific articles and health-related social media posts have been generated and shared daily, albeit with very limited use for ADE study and with little known about the content with respect to ADEs.

Objective: The aim of this study was to develop a big data analytics strategy that mines the content of scientific articles and health-related Web-based social media to detect and identify ADEs.

Methods: We analyzed the following two data sources: (1) biomedical articles and (2) health-related social media blog posts. We developed an intelligent and scalable text mining solution on big data infrastructures composed of Apache Spark, natural language processing, and machine learning. This was combined with an Elasticsearch No-SQL distributed database to explore and visualize ADEs.

Results: The accuracy, precision, recall, and area under receiver operating characteristic of the system were 92.7%, 93.6%, 93.0%, and 0.905, respectively, and showed better results in comparison with traditional approaches in the literature. This work not only detected and classified ADE sentences from big data biomedical literature but also scientifically visualized ADE interactions.

Conclusions: To the best of our knowledge, this work is the first to investigate a big data machine learning strategy for ADE discovery on massive datasets downloaded from PubMed Central and social media. This contribution illustrates possible capacities in big data biomedical text analysis using advanced computational methods with real-time update from new data published on a daily basis.

(JMIR Med Inform 2017;5(4):e51) doi: 10.2196/medinform.9170

KEYWORDS

adverse drug event; adverse drug reaction; drug side effects; machine learning; text mining

Introduction

Background

Adverse drug events (ADEs), defined as the set of detriments or injuries caused by a medication, have led to additional medical costs, prolonged hospitalization, morbidity, and ascribable disability worldwide [1-4]. ADEs encompass all adverse drug reactions but also include preventable causes of errors such as inappropriate dosing, dispensing errors, and drug abuse. Discovery of ADEs has gained great attention in the health care community, and in the last few years, several drug risk-benefit assessment strategies have been developed to analyze drug efficacy and safety using different medical data sources, ranging from electronic health records (EHRs) to human-health–related social media and drug reviews [5-14]. A variety of combined computational methods using natural language processing (NLP), machine learning strategies, and text retrieval algorithms have been employed to extract ADEs from such data sources [15-23]. Clinical trials, EHRs, and medical case reports are additional biomedical data-rich sources that have been utilized for ADE extraction [24-28].

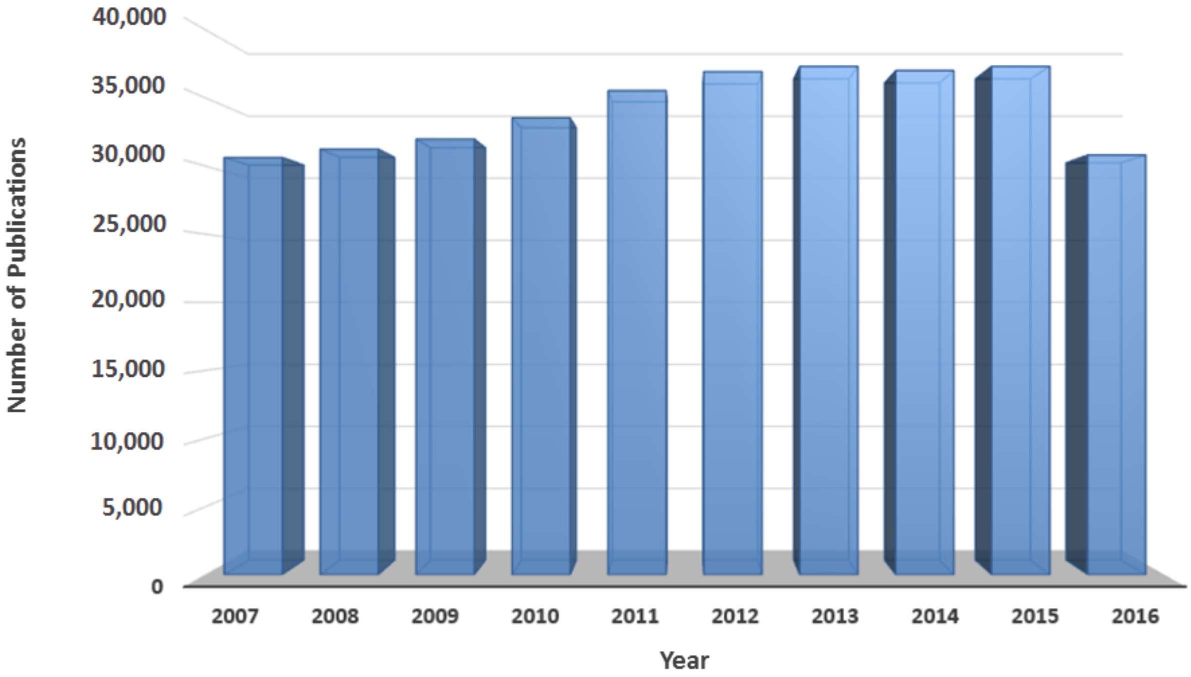
In recent years, biomedical articles produced by scientists all across the world have grown extensively. Figure 1 [29] shows that the number of journal and conference papers published in different medication studies (eg, ADEs and drug analysis, drug evaluation, and drug repositioning) rapidly grew in number from year 2007 to 2016. The total number of publications in those years is approximately 342,301 articles. To roughly estimate the size of such scientific papers, we assumed a PDF file format for each article. The size of a PDF file depends on the number of pages and pictures or metadata inside the file. Considering a 9 to 11 page PDF including plain text along with a few pictures, it may equal almost 3 MB size in average, and it appears, approximately 1.02 TB articles were generated in drug associated studies from 2007 to 2016. The other file formats such as extensible markup language (XML), may be much larger in size. Scientific articles published in biomedical research are usually generated using standardized and principled methods and therefore, are especially valuable for high-quality knowledge discovery. This great deluge of information includes an enormous number of scientific publications on ADEs' study, an area of focus into which many biomedical researchers have entered, developing a variety of research activities for discovering, analyzing, and monitoring ADEs [30-38].

It is impossible for researchers, scientists, and physicians to read and process the large body of scientific articles and remain abreast of the foremost information regarding ADEs. Therefore, there is a pressing need to develop intelligent computational methods, particularly big data analytics solutions, to efficiently process this wealth of data. Big data biomedical text analysis utilizes advanced computational technologies including big data infrastructure, NLP, statistical analytics, and machine learning algorithms to extract facts from text data. This in turn generates new hypotheses by systematically analyzing large numbers of scientific publications.

Objectives and the Main Contributions

Whereas ADE discovery from diverse biomedical data sources in general has been studied historically in health care informatics, the use of big data scientific articles and health-related social media for ADE discovery has been very limited so far. The motivation of this work is to study big data machine learning solutions, particularly big data neural networks (bigNN), to analyze ADEs from large-scale biomedical text data, developing a scalable framework to fulfill the following objectives: (1) to extract current knowledge and high-quality information about ADEs using full text scientific articles and social media, (2) to utilize and adapt advanced NLP and machine learning algorithms in a large-scale fashion by the use of big data infrastructures, and (3) to provide better insights and tendencies in large-scale biomedical text analytics and identify the challenges and potential enhancements toward efficient and accurate ADE discovery. We briefly summarize our *main contributions* as follows:

Figure 1. The number of publications in several medication studies available at PubMed over the last 10 years. The results obtained by submitting a query: ((((((((((drug analysis[MeSH Terms]) OR drug analysis[MeSH Subheading]) OR adverse drug event[MeSH Terms]) OR adverse drug event[MeSH Subheading]) OR adverse drug reaction[MeSH Terms]) OR adverse drug reaction[MeSH Subheading]) OR drug evaluation[MeSH Terms]) OR drug evaluation[MeSH Subheading])) OR drug repositioning[MeSH Terms]) OR drug repositioning[MeSH Subheading].



We initiated a study of big data literature mining for ADE discovery with the use of two different data sources: (1) published full text scientific articles available on PubMed Central [39], and (2) posts available in health-related social media, including MedHelp [40], patient [41], and WebMD [42]. Although several promising approaches have been designed for biomedical text mining, the development of scalable machine learning frameworks capable of ADE extraction from big data is very limited. To the best of our knowledge, our work is the first to investigate a bigNN strategy for ADE discovery on massive datasets downloaded from PubMed Central and social media.

- With the current work and using big data analytics platforms such as Elasticsearch and Apache Spark, we developed a scalable framework to analyze and visualize ADEs from hundreds of thousands of published scientific articles and social media blog posts.
- Combining a variety of the internal neural network parameters, we presented a predictive model that obtained accuracy, precision, recall, and area under 92.7%, 93.6%, 93.0%, and 0.905, respectively, on a massive dataset downloaded from PubMed Central plus health-related social media.
- This paper opens the door to pursue large-scale biomedical literature mining and its application in health care informatics in general and introduces several possible enhancements to advance the level of the impact of this research area.

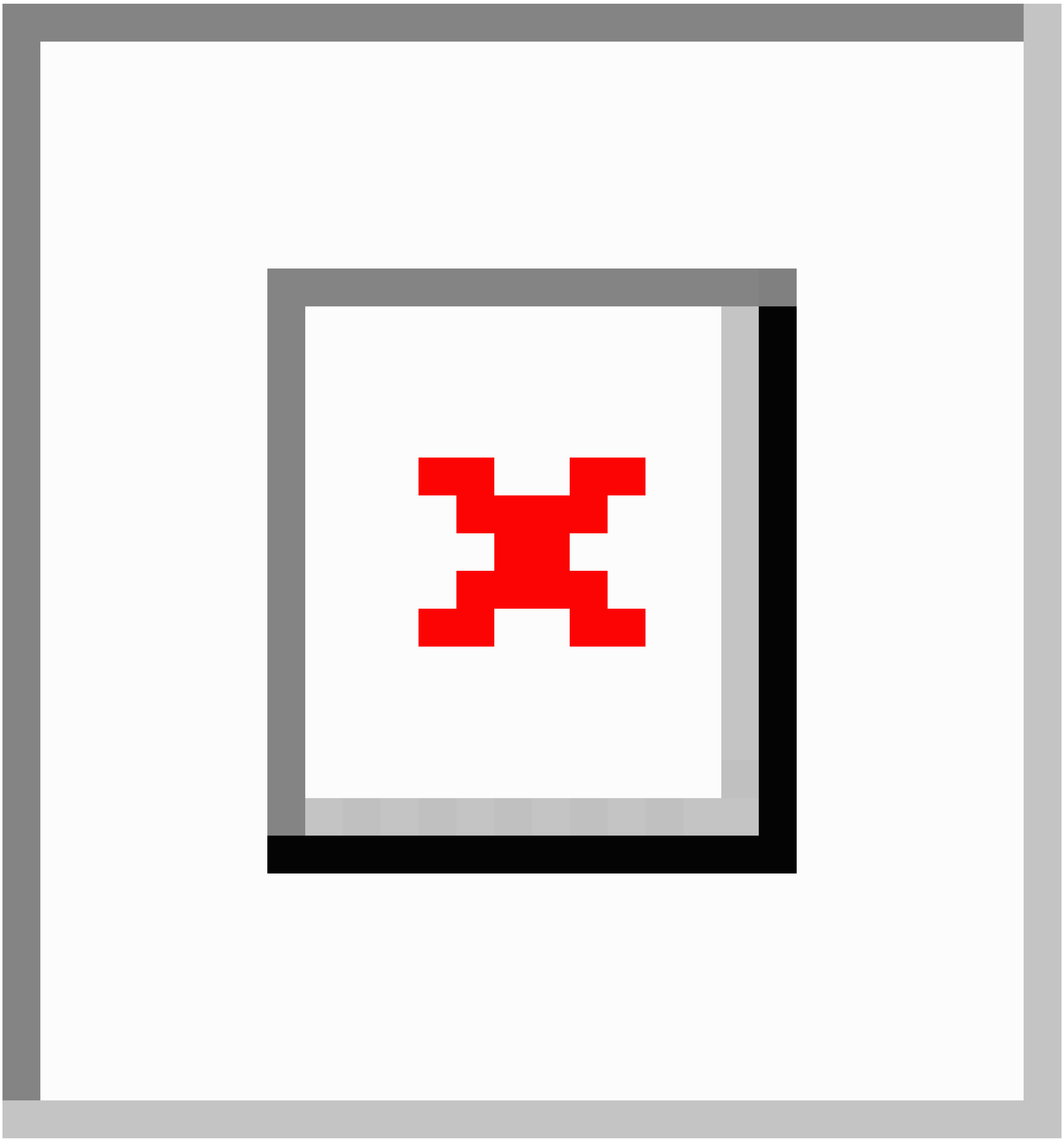
Methods

The general pipeline of the proposed ADE extraction framework is illustrated in Figure 2. In this section, we shall explain the underlying tiers of the proposed bigNN framework.

Tier 1: Data Access

Tier 1 systematically collects the expanding body of scientific articles and social media blog posts through different data sources available on the Internet. A multi-threaded crawler or downloader was developed to provide timely and efficient processing of the diverse big data content found on the Internet. *Scrapy* [43], a free and open source Web crawling system has been used to allow multiple threads to automatically fetch URLs from different sources. A queuing system and a scheduler have also been established as a part of the data access tier.

Figure 2. The proposed system for adverse drug event (ADE) discovery. All tiers developed on top of the Apache Spark 2.0 that utilizes an Elasticsearch database 2.4.1 to data storage and retrieval.



All collected data (eg, XML files) are first turned into plain text data, and together with associated metadata (eg, journal name, author list, and publication date) are stored in a single type (table type) inside a No-SQL database, namely Elasticsearch [44], which provides a distributed, open source, RESTful and full text search and analytics engine.

Tier 2: Natural Language Processing

Tier 2 includes several computational procedures to process raw text data, preparing potential ADE sentences to feed the next tier.

Selecting Relevant Documents

One of the major goals of the proposed system was to collect groups of sentences that provide evidence about drug-event pairs. Toward that outcome, there is an emergent need to identify trustworthy and reputable data sources (eg, well-founded and prestigious journals such as *Nature*, *PNAS*, and *PLOS*). As the proposed framework accumulates data from two separate data sources, including scientific journals and messages posted on social media, we established two different criteria to yield more credible data. Section A.1 of Multimedia Appendix 1 further discusses the proposed method and criteria.

Text Processing

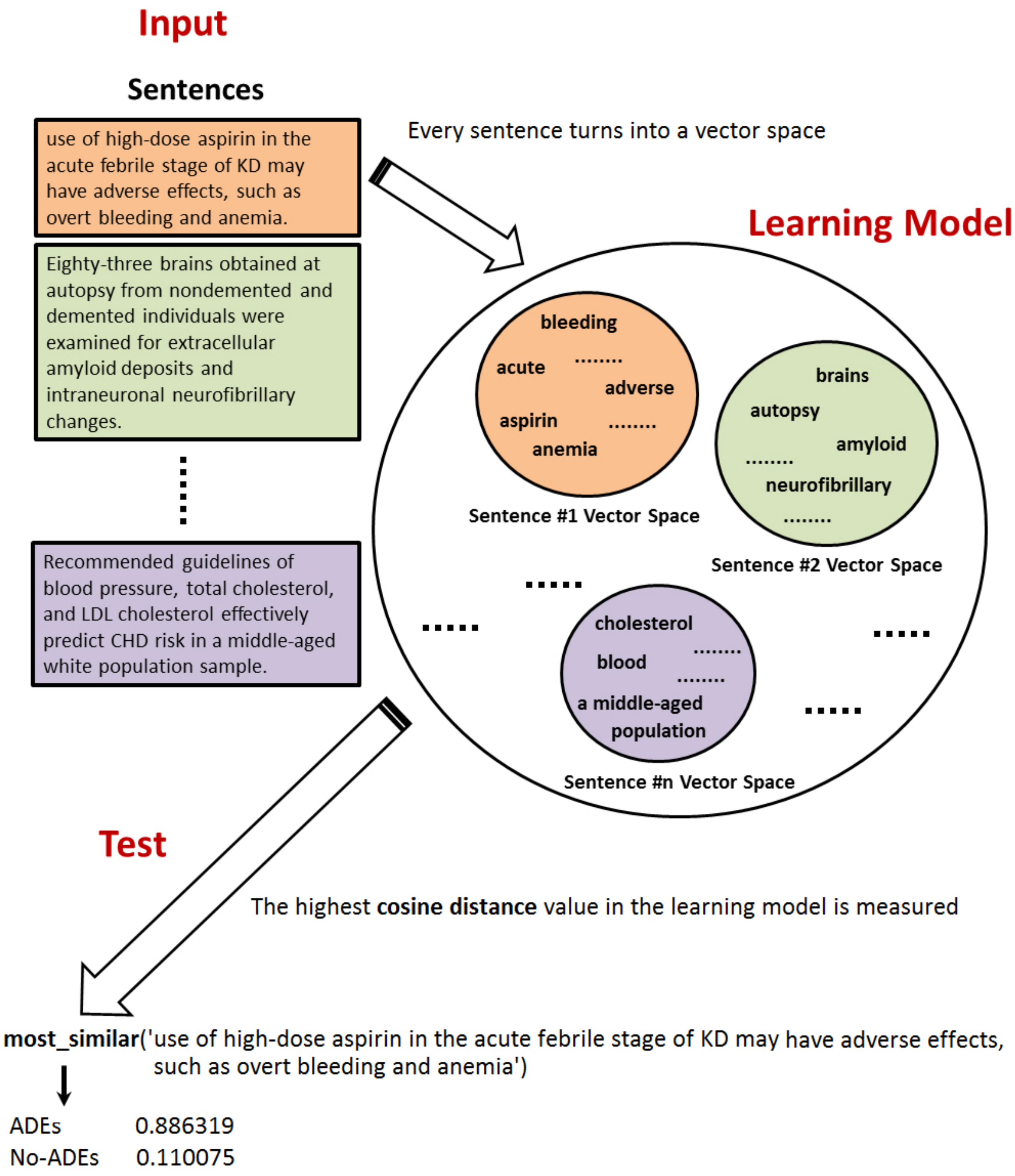
We first normalize all documents by converting corpora into a standard consistent form. This process (1) converts all characters to lower case, (2) transliterates to American Standard Code for Information Interchange if needed, and (3) deletes a set of existing substrings and patterns (eg, [], , ?, !, and ()). Once we complete the proposed text normalization process, we convert every document into a set of sentences. Although several ADEs could be captured among different sentences, extracting ADEs interactions across sentences is significantly more challenging than within sentences [6,45]. To feed the bigNN system, a random subset of sentences was selected for manual annotation by three domain experts; see Section A.2 of Multimedia Appendix 1 for details. The random subset of sentences includes health and medical-related text data either with or without ADE interactions. Sentences that are missing either a drug name or an adverse drug effect term were excluded. Using a Web application (Multimedia Appendix 1), the domain experts labeled individual sentences as ADEs, No-ADEs, or Not Decided. To focus on binary classification, we omitted sentences labeled “Not Decided” leaving two different classes: (1) ADEs and (2) No-ADEs. Section A.2 of Multimedia Appendix 1 explains how we made a training set for the machine learning tier.

Tier 3: Machine Learning

Tier 3 implements the core functionality needed for making a predictive model to distinguish ADEs sentences versus those that are No-ADEs. From the machine learning perspective, this means making a binary classification predictive model that assigns one of two trained labels to new unlabeled sentences. Therefore, when an arbitrary sentence arrives, the predictive model chooses one category exclusively from two predefined classes as ADEs or No-ADEs. There are two basic steps. The first step is to extract effective content as a set of features. The next step is the text classification assignment. The bag-of-words (BoW) representation, a widely used content extractor algorithm, has been around for several years in the text analytics domain, and it provides an easy way to turn text-based data records into a set of feature vectors such that the frequency of occurrence of words (eg, uni-grams and/or bi-grams, along with part-of-speech [POS] tagging) in the corpus is used as a feature vector to train a classifier (eg, support vector machine [SVM], decision tree, and/or logistic regression) [46-51]. The BoW representation maintains word intensities across the corpus, but it dissembles grammar, syntactic, semantic, and word order. In contrast to the BoW representation, the word2vector (word2vec) algorithm, originally developed at Google [52], includes a set of computational methods that turns a corpus of text data into a meaningful vector space that encompasses grammar, semantic, and word order. Word2vec comes with a two-layer neural network and is able to tackle several text analytics functions including dependency parsing [53,54], named entity recognition [55,56], text classification [57,58], and word clustering [59]. The model takes a text corpus as an input and turns each word in the corpus into a vector as illustrated in Section A.3 of Multimedia Appendix 1. It then groups vectors of similar words together in a vector space, training words against other words that neighbor them in the input text corpus [60,61]. Our bigNN system implemented the word2vec neural network, which is fully explained in Section A.3 of Multimedia Appendix 1.

An abstract view of the proposed learning algorithm is shown in Figure 3. As mentioned in Section A.3 of Multimedia Appendix 1, the algorithm builds a vector representation for words and/or sentences. If we train a learning model with 98-dimension, then we will obtain a 98-digit in front of each word (eg, “aspirin”). The cosine distance similarity [62,63], which is the normalized dot product between vectors, is then measured to find the best fitness class for a sentence. Once we have the labeled sentences as ADEs or No-ADEs, we focus on ADE sentences and find the positive adverse-drug interactions using cTAKES [64].

Figure 3. This figure depicts the abstract view of the proposed learning model where the cosine distance similarity is measured to select an appropriate category for a given sentence. Every single word is represented as a vector, and eventually every sentence (eg, adverse drug events [ADEs] or No-ADEs) turns in to a vector space. Once we have done with training the model, for every new coming sentence, the highest cosine distance value is then measured to find the best fitness class.


Input
Sentences
use of high-dose aspirin in the acute febrile stage of KD may have adverse effects, such as overt bleeding and anemia.
Eighty-three brains obtained at autopsy from nondemented and demented individuals were examined for extracellular amyloid deposits and intraneuronal neurofibrillary changes.
Recommended guidelines of blood pressure, total cholesterol, and LDL cholesterol effectively predict CHD risk in a middle-aged white population sample.
Every sentence turns into a vector space
Learning Model
bleeding
acute
....
adverse
aspirin
....
anemia
Sentence #1 Vector Space
brains
autopsy
amyloid
....
neurofibrillary
....
Sentence #2 Vector Space
cholesterol
blood
....
....
a middle-aged
population
Sentence #n Vector Space
Test
The highest cosine distance value in the learning model is measured
most_similar('use of high-dose aspirin in the acute febrile stage of KD may have adverse effects, such as overt bleeding and anemia')
ADEs 0.886319
No-ADEs 0.110075

Results

Implementation and Test Bed

We investigated two different data sources: (1) biomedical articles and (2) health-related social media blog posts. The first data source included 97,246,719 sentences obtained from almost 1,451,413 abstracts and full text articles available on PubMed Central, and the second one consisted of 2,524,622 sentences obtained from 419,915 blog posts at MedHelp, Patient, and WebMD. To train our proposed predictive model illustrated in Figure 2, we first randomly selected different subsets for the purpose of manual annotation. Those subsets were annotated by three domain experts who have been working in medical and pharmacy domains. We defined two different classes as ADEs and No-ADEs to indicate drug-event interactions in a sentence. The sentences that included positive drug-event interaction were tagged as ADEs, and the others were tagged as No-ADEs. Before the large-scale manual extraction, we evaluated the interrater reliability among our three domain experts, the results of kappa statistics, .84, which indicates very good concordance and agreement; more details are reported in Section A.2 of Multimedia Appendix 1.

Table 1. Nine datasets were employed to make and evaluate the big data neutral network (bigNN) system. Each row identifies the dataset along with the number of ADEs and No-ADEs sentences within the dataset. Every row also shows how many of these sentences are human-labeled and/or machine-labeled. The datasets are separated into three main categories: biomedical articles (*_BA), social media posts (*_SM), and the combination of the two (*_Combined).

Dataset ID	Total number of sentences	Number of ADEs[a] sentences	Number of No-ADEs sentences	Number of human-labeled sentences	Number of machine-labeled sentences
ADEs#1_BA	6960	3311	3649	6960	0
ADEs#1_SM	400	160	240	400	0
ADEs#1_Combined	7360	3471	3889	7360	0
ADEs#2_BA	13,545	6359	7186	7015	6530
ADEs#2_SM	472	195	277	405	67
ADEs#2_Combined	14,017	6554	7463	7420	6597
ADEs#3_BA	21,278	10,307	10,971	7015	14,263
ADEs#3_SM	565	241	324	405	160
ADEs#3_Combined	21,843	10,548	11,295	7420	14,423

[a] ADEs: adverse drug events.

Next, we developed nine different datasets from those sentences to train the proposed bigNN system. The first set of datasets (the first three rows in Table 1) used only human-labeled datasets from our domain experts. Within these three datasets, three types of sentences were generated: the first type was biomedical articles, the second was social media, and the third was the combination of the two. For the second set of datasets (the three datasets in the middle of Table 1), we added some machine-labeled datasets (the number of datasets were reported in the last column, “number of machine-labeled sentences” of Table 1) with human-labeled datasets to increase the sample sizes. We also generated three types of sentences as illustrated above. For the third set of datasets (the last three rows of Table 1), we added in more machine-labeled datasets with human-labeled datasets; similar three types of sentences were generated.

We also utilized three smaller datasets, including 2600, 3500, and 4000 human-labeled data records and plotted learning curves to see how the accuracy of the predictive model varies with increasing amount of training data. The results were not promising enough, and we started with a larger human-labeled dataset as illustrated in the first three rows in Table 1. For all the experiment, we utilized 75% of every dataset to train the model and 25% to test it using four-fold cross validation, with no sentence to appear in both the training and testing sets at the same time.

Experimental Setup

Every programming module in Tier 1 and Tier 2 was developed by Python 2.7.13. Tier 3, the machine learning tier, was implemented by Java j2SE 8. All of these tiers were developed on top of the Apache Spark 2.0 [65] and Elasticsearch DB 2.4.1 [44] just to tackle the problem of big data analytics in an efficient and timely fashion. From the computational side, a dedicated computational resource, including two virtual machines in a VMWARE cluster environment, each running a 64-bit CentOS 6.8 operating system with 8 vCPUs, 16 GB RAM, and 1 TB HDD in total, hosted on a Xeon E5-2690V3 2.6 GHz CPU, were used to obtain the experimental results.

Experimental Validations

We analyzed the performance of the predictive model across all the datasets. Accuracy, precision, and recall obtained by the experiments are shown in Multimedia Appendix 2. The first column shows the dataset used to in the experiments. The second column describes a configuration setup for a set of internal parameters of the proposed neural network model. Minimum word frequency (MWF) allows for ignoring all words in the vocabulary with total occurrences lower than MWF value. Epoch (EP) is the number of forward and backward passes of all training examples. Window size (WS) defines context windows size to generate a vector representation for words across the documents. Iteration (ITR) defines the number of iterations done for each mini-batch during a training process. The last column shows elapsed time for the training stage. This does not reflect the time of text-preprocessing tasks such as normalization and tokenization. The current table shows that greater EP and ITR with the use of WS of two will provide better performance across all three datasets. To further analyze our proposed ADEs sentence discovery system, we also compared the proposed predictive model with the combination of BoW feature selection method and SVM, decision tree, and naïve Bayes classifiers. Uni-grams, bi-grams, along with POS tagging were used as BoW features to make a predictive model. One can see in Multimedia Appendix 2, the most promising accuracy results across all datasets obtained by the (MWF=2%, EP=25, WS=2, and ITR=10) configuration. Furthermore, we demonstrated that the performance of bigNN system, both accuracy and time of completing the task, is comparable with traditional SVM, naïve Bayesian, and decision tree with BoW strategy. The results of this experiment are shown in Table 2. All the measures in Table 2 are selected from the best performed model by tuning the models using different parameters for all the bigNN system and SVM, naïve Bayesian, and decision tree with BoW strategy. Regarding the BoW feature set, we obtained the best results by utilizing a combination of uni-grams and bi-grams, together with POS tagging. With respect to the traditional machine learning classifiers, and for example SVM, the best performance was achieved with the use of radial basis function kernel, loss of 0.12, seed of 1, and without normalization of input data.

For each of the datasets shown in Table 2, we split the data randomly as 75% to train and 25% to test the proposed sentence classifier system. The best accuracy results using our proposed predictive model were obtained by MWF=2%, EP=25, WS=2, ITR=10 configuration. Regarding the BoW features set, we utilized a combination of uni-grams, bi-grams, and POS tagging. Using the proposed predictive model, the vocabulary size for ADEs#1_Combined, ADEs#2_Combined, and ADEs#3_Combined datasets were 15125, 26,448, and 37,524, respectively.

Whereas with the use of BoW, when it utilized uni-grams, bi-grams, and POS, the vocabulary size was 33,567, 59,941, and 76,758 for datasets ADEs#1_Combined, ADEs#2_Combined, and ADEs#3_Combined. For SVM, decision tree, and naïve Bayes classifiers, we utilized Weka library (version 3.7.12) [66] running on hadoop-2.7 [67] by the use of Hadoop distributed file system.

The study shows that bigNN system generates better results in comparison with traditional BoW along with SVM, decision tree, and naïve Bayes classification algorithms. Area under the curve (AUC) of our proposed predictive model across all three datasets was also analyzed, and it is shown in Figure 4.

The current test results are only as good as the predictive model developed in the training phase. We accomplished further experiments just to make sure that the predictive model is sufficiently accurate in assigning appropriate classes for new unlabeled data records. Our approach was to rely on human reviews. We fed the proposed word2vec predictive model with new unlabeled data records and gave a random subset of the system output five hundred system-labeled instances to two domain experts to review. We got 87.6%, 86.1%, and 88.7% in average for accuracy, precision, and recall, respectively.

We briefly summarized the experimental results as follows:

The results we obtained showed that the use of combined dataset was better than the use of either source individually (Table 1) for all the models and is statistically significant (at *P*=.04).

The results illustrated in the Multimedia Appendix 2 show that a greater *epoch* along with a greater *iteration* with the use of *window size* of two tend to be useful over all datasets using bigNN system, and the result is statistically significant (at *P*=.02). However, it requires a longer training time.

The comparative study shown in Table 2 demonstrates that the bigNN system was able to generate better results in comparison with traditional BoW along with SVM, decision tree, and naïve Bayes classification algorithms. Performing a *t* test on AUC matched by those models shows statistically significant differences (at *P*=.03) between our bigNN system and those two models utilizing BoW along with decision tree and naïve Bayes. It also shows no statistically differences employing BoW and SVM.

Table 2. The comparisons of our big data neutral network (bigNN) system with traditional bag-of-words (BoW) method using support vector machine (SVM), decision tree, and naïve Bayes classifiers.

Dataset ID	Learning method	Number of sentences	Accuracy (%)	Precision (%)	Recall (%)	Area under the receiver operating characteristic	Training time (min)
ADEs#1_Combined[a]	bigNN[b] system	7360	88.7	88.5	89.4	0.842	45.7
ADEs#1_Combined	BoW[c] + SVM[d]	7360	89.4	88.3	88.0	0.841	66.3
ADEs#1_Combined	BoW + decision tree	7360	84.0	83.7	82.1	0.775	49.5
ADEs#1_Combined	BoW + naïve Bayes	7360	83.7	82.1	83.5	0.763	48.9
ADEs#2_Combined	bigNN system	14,017	89.1	88.9	89.3	0.874	69.5
ADEs#2_Combined	BoW + SVM	14,017	89.5	88.0	89.7	0.875	88.9
ADEs#2_Combined	BoW + decision tree	14,017	85.5	84.9	84.5	0.861	75.2
ADEs#2_Combined	BoW + naïve Bayes	14,017	84.3	84.0	85.7	0.855	73.8
ADEs#3_Combined	bigNN system	21,843	92.7	93.6	93.0	0.905	121.7
ADEs#3_Combined	BoW + SVM	21,843	92.5	94.0	93.2	0.911	159.5
ADEs#3_Combined	BoW + decision tree	21,843	88.3	87.5	87.2	0.868	131.5
ADEs#3_Combined	BoW + naïve Bayes	21,843	87.5	86.2	85.8	0.851	135.3

[a] ADEs: adverse drug events.

[b] bigNN: big data neutral network.

[c] BoW: bag-of-words.

[d] SVM: support vector machine.

Figure 4. This figure shows the area under the curve (AUC) of our proposed predictive model. ADEs: adverse drug events.

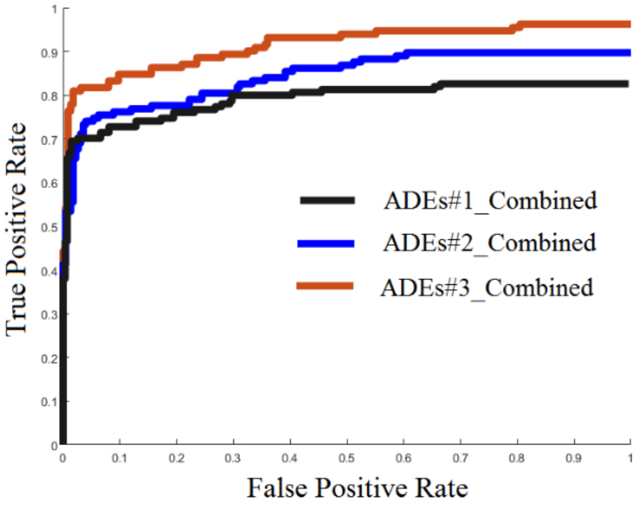


Table 2 also shows that our bigNN system was faster than the BoW method along with SVM, decision tree, and naïve Bayes classification algorithms. Performing a *t* test on training time matched by those methods presents statistically significant differences (at *P*=.03) between our proposed model and all those three models developed by BoW along with SVM, decision tree, and naïve Bayes.

The AUCs presented in Figure 4 show that larger training samples tend to add benefit when making an accurate and reliable predictive model.

To further analyze the results, the next section will present a set of ADEs' scientific visualizations obtained using the proposed framework.

Scientific Visualization of ADEs

Scientific visualization is concerned with representing large and highly dimensional information by means of charts, graphs, and images. The general objective of any scientific visualization is to improve understanding of the data being investigated. In this section, we scientifically visualize the ADE information extracted by our proposed system. Once the system was trained across the different datasets, we fed the system new unlabeled data. This included 92,681,359 sentences from biomedical articles downloaded from PubMed Central and 1,624,117 sentences from social media, including MedHelp, Patient, and WebMD. There are considerable amounts of drugs and adverse events, and it is beyond the scope of this paper to visualize all of them. Here, we limited ourselves to the list of 28 drugs as illustrated in Table 3. The proposed system could find 12,265 ADE sentences from biomedical articles and 181 ADE sentences from three social media sources using the drug list defined.

Figures 5 and 6 show ADE discovery visualization results obtained from the biomedical articles and social media, respectively. Figure 7 represents a set of word cloud examples generated for ADEs associated with "aspirin," "atenolol," "gabapentin," and "statins." A word cloud is a graphical representation composed of different words contained in a corpus, in which the size of every word indicates its frequency or importance to the text.

Table 3. The list of the drugs used to make the scientific visualization results.

Row ID	Drug name
1	Anesthesia
2	Antihistamine
3	Antipsychotic
4	Aspirin
5	Atenolol
6	Atorvastatin
7	Azithromycin
8	Dexamethasone
9	Diazepam
10	Dopamine
11	Ephedrine
12	Gabapentin
13	Galantamine
14	Heparin
15	Ibuprofen
16	Lamotrigine
17	Lorazepam
18	Melatonin
19	Meloxicam
20	Metformin
21	Methylphenidate
22	Ondansetron
23	Orlistat
24	Sildenafil
25	Statins
26	Vioxx
27	Warfarin
28	Wellbutrin

Figure 5. The adverse drug events' (ADEs') visualization results obtained from biomedical articles. One can see the number of "anesthesia" observations is 2186, where the most frequent adverse drug events are "hypotension," "nausea," "aspiration," and "depression," respectively. Using the CI of 95%, Pr(hypotension|anesthesia) is between 20.5% and 23.5%, and Pr(nausea|anesthesia) is some point between 12.0% and 14.7%. Another example is "gabapentin" where the most frequent adverse drug events based on ADE sentences extracted from biomedical articles are "dizziness," "nausea," "fatigue," and "edema" whereas the number of "gabapentin" is 261. Using the CI of 95%, Pr(dizziness|gabapentin) is between 33.0% and 44.2%.

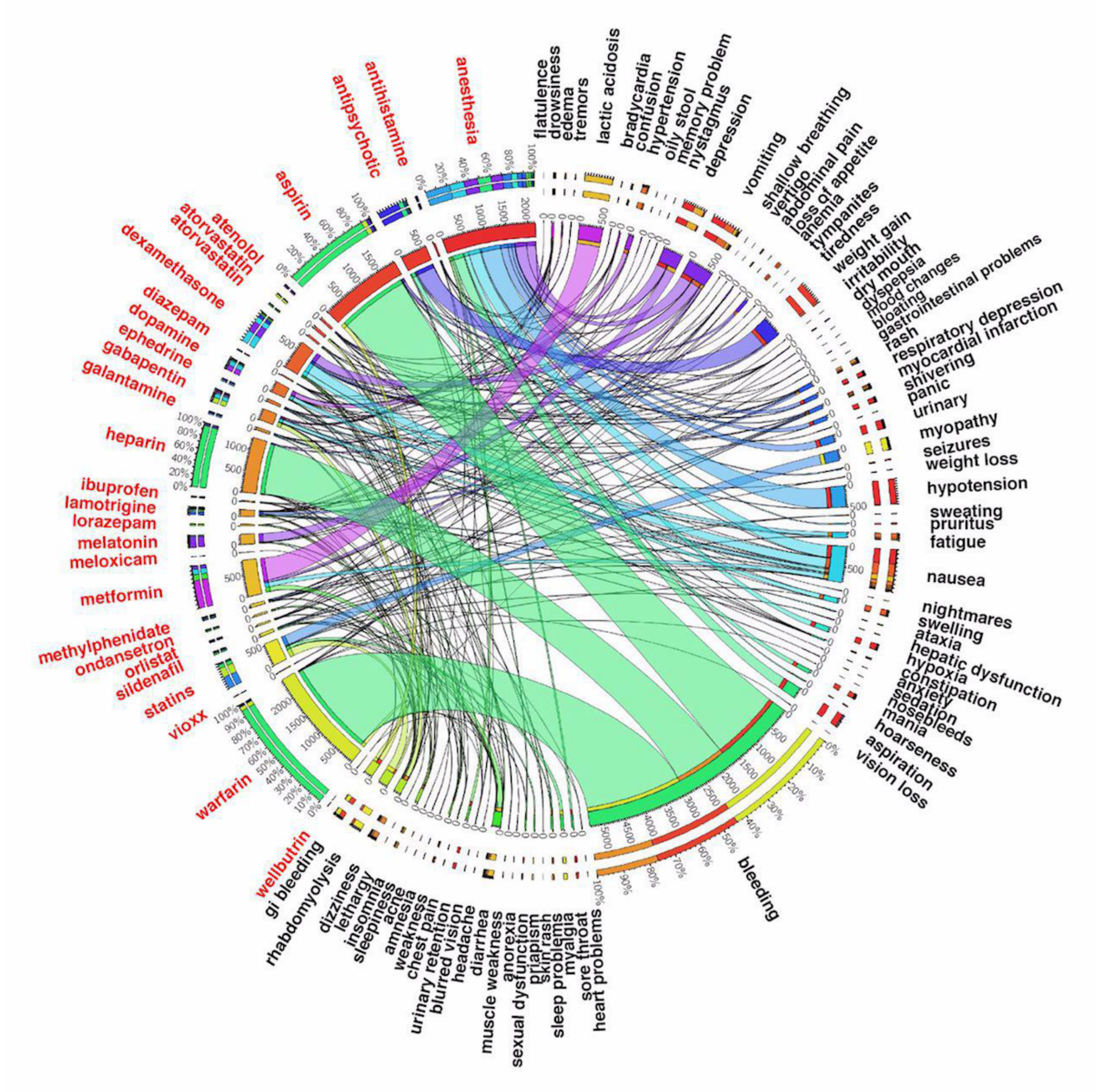


Figure 6. The adverse drug events' (ADEs') visualization results obtained from health-related social media. One can see the number of "metformin" observations is 28, where its most frequent ADEs are "nausea," "diarrhea," "vomiting," "dizziness," and "stomach pain," respectively. Another example is "atenolol" and its most frequent adverse events are "depression," "bradycardia," "hypotension," "tiredness," and "dizziness." From the adverse events aspect, this figure shows, for example, "nausea" as an adverse event is mostly associated with "metformin," "dexamethasone," "antihistamine," "wellbutrin," and "sildenafil".

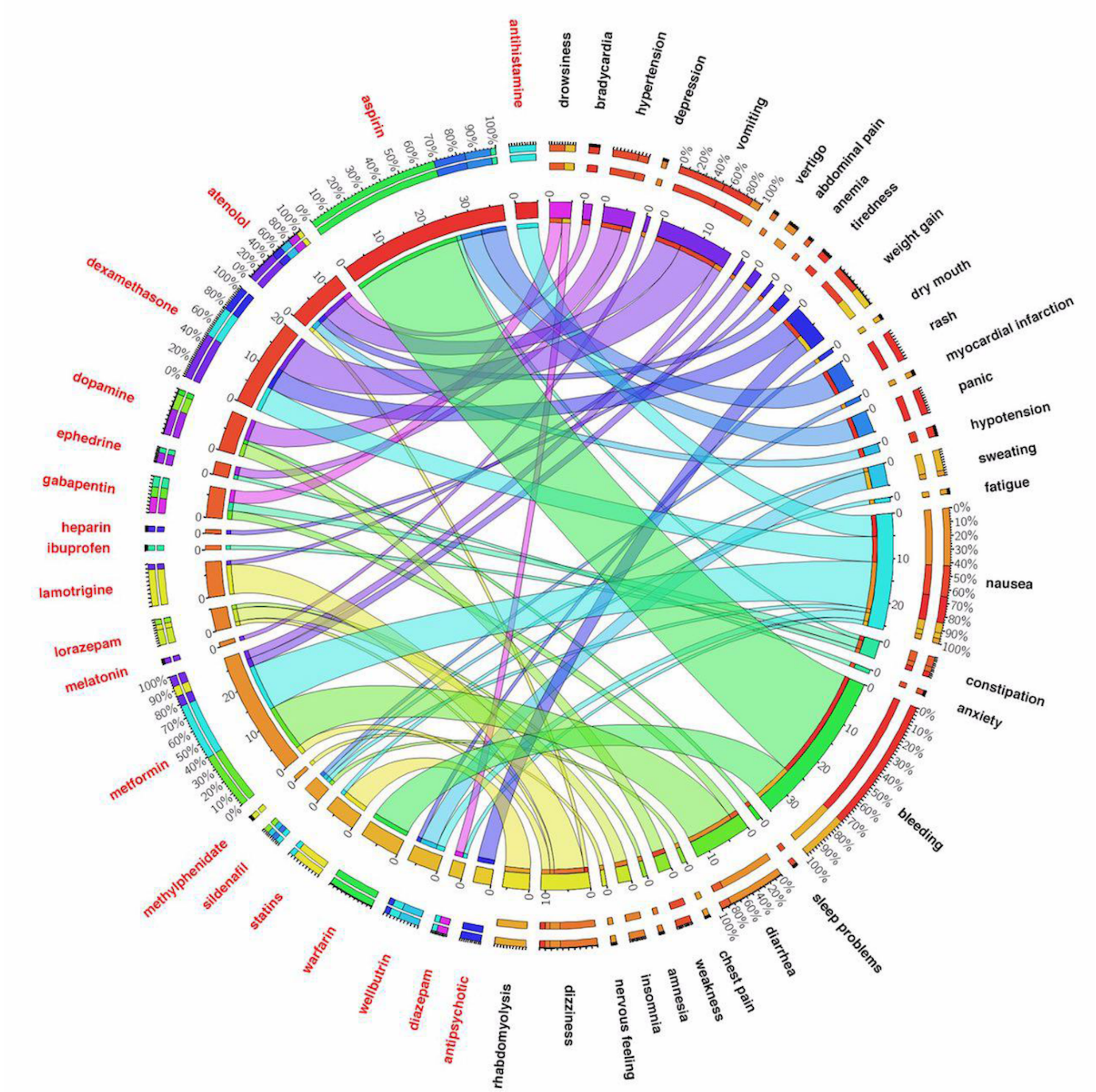
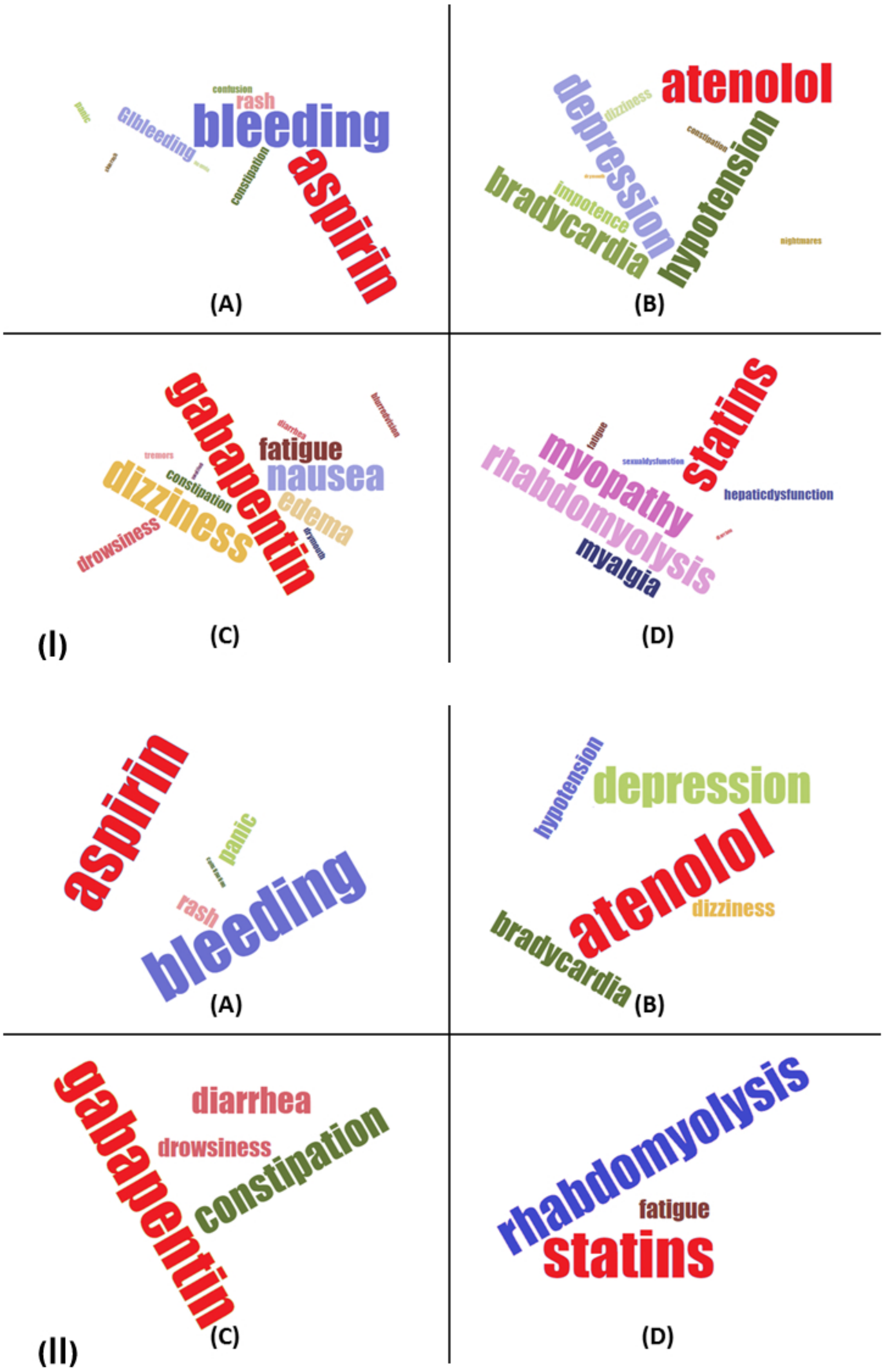


Figure 7. The word cloud representations for “aspirin,” “atenolol,” “gabapentin,” and “statins.” (1) These results are obtained from biomedical articles. Whereas the most frequent adverse dug events for “statins” are “myopathy,” “rhabdomyolysis,” “myalgia,” “fatigue,” and “hepatic dysfunction” as shown in (D), the most frequent ADEs for “atenolol” are “depression,” “hypotension,” “bradycardia,” and “impotence,” respectively (B). (2) These results are extracted from social media. Whereas the most frequent ADEs for “statins” are “rhabdomyolysis” and “fatigue” (D), the most frequent ADEs for “atenolol” are “depression,” “bradycardia,” “hypotension,” and “dizziness,” respectively (B).



Discussion

Principal Findings

Modern medical data sources ranging from clinical trials, EHRs, and medical case reports to scientific articles and patients’ blog posts are rapidly growing in size and complexity, and scientific biomedical articles, as well as health-related social media are under-researched data sources for biomedical studies. Thus, there is a pressing need to develop efficient solutions to harness this wealth of data using advanced computational methods such as artificial intelligence and big data machine learning.

With this contribution, an attempt was made to design and develop a computational bigNN system to detect, analyze, and visualize ADEs from massive data sources obtained from PubMed Central, a widely referenced repository of scientific articles, and the social media blog posts existing in MedHelp, Patient, and WebMD. The model was developed using the word2vec neural network architectural on top of the Apache Hadoop cluster, Apache Spark, and Elasticsearch No-SQL distributed database to tackle efficient big data ADE identification. We accomplished extensive experimental validations to ensure that the proposed predictive model can accurately assign appropriate classes (eg, ADEs or No-ADEs) for both current and also new unlabeled data records. Our trained system was able to detect a number of well-known ADEs from unlabeled data taken directly from the literature. A list of well-known ADEs can be found in Multimedia Appendix 1. Warfarin is an extremely effective anticlotting agent primarily used for patients at high risk for stroke or heart attack because of atrial fibrillation that, up until recently, has been a cornerstone of treatment. One common and potentially serious adverse effect of warfarin therapy is bleeding. This can occur because of changes in diet, drug interactions, or spurious physiological changes. The effectiveness of warfarin coupled with a high risk for serious bleeding events led to extensive research and publication in the medical literature, which is very apparent in our data visualization. Similarly, the widespread use of aspirin, which also carries a risk of bleeding, has been extensively studied for primary prevention of heart attack, colorectal cancer, and for secondary preventions of cardiovascular events, which is also captured in our results. Our system also identifies serious but rare side effects such as lactic acidosis caused by metformin use and rhabdomyolysis attributed to statin therapy. Common drug side effects are also captured, although with a smaller number of hits. Examples in Figure 5 include ADE pairs of metformin and diarrhea, antipsychotics and weight gain, gabapentin and dizziness, and drowsiness with antihistamines.

The findings of our system shed light on areas for future work and on inherent challenges with semantics and context in NLP. The following examples will illustrate some of these challenges. Our system identifies nausea and vomiting as ADEs associated with dexamethasone. Although it is true that dexamethasone can cause these reactions, it is commonly prescribed to prevent nausea and vomiting associated with chemotherapy. Without more contextual clues from the free text in the articles, it is impossible for us to decide whether dexamethasone is being identified as a treatment or a causal agent in this case. Another example where our system identified an indication as an ADE was with lamotrigine and seizures. Seizure is a primary indication for the use of lamotrigine and not a causal agent. Successfully classifying these edge cases may require additional labeled data, a larger window size, or other unexplored techniques and is an area for further study.

The ADEs identified from our analysis of social media provide a number of interesting similarities and differences with those in the literature. In our social media results, as might be expected, we see a larger proportion of ADEs related to the more common side effects of drugs as compared with the literature. For example, we see a large proportion of sentences identified for nausea and diarrhea with the use of metformin and fewer mentions of abdominal pain and vomiting. This parallels nicely with the incidence expected in real-world use. In contrast, we see a high proportion of sentences labeled for lactic acidosis, an extremely rare ADE associated with metformin use from the literature. The number of sentences describing adverse drug events in biomedical text articles is highly variable and includes influencing factors such as the severity of the drug reaction, safety concerns eliciting directed study, and the goals and intent of the research paper. Non-life-threatening ADEs are less important to clinical researchers, assuming they do not result in discontinuation, compared with serious reactions. In a similar way, side effects reported in social media will naturally include more common side effects, particularly because they are impactful to the patient taking the medication. In future work, text mining should take advantage of these naturally occurring differences. Publications in the biomedical literature or postings in social media, especially early after the release of a novel drug, may include case studies or reports of side effects not seen in clinical trials that could be detected by our system before the signal reaches the critical detection threshold of reporting systems such as the FDA Adverse Event Reporting System.

Limitations

We acknowledge some limitations to this research study. Assessing the quality of scientific journals is a difficult task but important for narrowing the search space of candidate articles. In this work, we attempt to combine three ranking indices in the hope of identifying journals with the most credible information without generating a hand-curated list. It may be the case that our approach excludes journals that would be extremely useful in identifying ADEs but are excluded based on a low combined score. In addition, some journals will provide a richer source of information on ADEs than others based on their intended audience and subject matter irrespective of any ranking criterion. This leaves the question of which journals to focus on for ADE text mining open for further exploration. In the scientific articles, as well as social media blog posts, we noticed that different people may use different terms to discuss a similar single adverse event. For example, the terms “mood changes” and “mood swings” are often used interchangeably, equally meaning “mood changes.” A robust dictionary-based methodology may help address this issue. Additionally, text mining of the social media comments posted by patients is a really challenging task, as the comments are often written in an informal way. As we have a smaller number of labeled sentences from this source, we didn’t address overcoming the differences in phrasing, spelling errors, or other problems introduces from these posts. No fuzzy matching or specialized dictionaries were used on either source, so if there are spelling errors in the drug name, adverse event, or indication, the sentence would have been excluded from evaluation. Furthermore, our proposed big data neural network model is more appropriate for the short-length text data (eg, a sentence) classification rather than the long-length text data (eg, full text articles) categorization. Advanced tokenization systems, and in particular, a medical literature–based tokenization system will be useful with short-length text data.

Conclusions

The present contribution utilized a bigNN system to discover only ADEs; however, there are several interesting applications to leverage the proposed system. For example, the proposed big data analytics pipeline could help study drug repurposing and impact drug development, particularly in analyzing the success rates of new medications. The social media demographic information (eg, age, gender, ethnicity, and location) supports use of the current contribution to explore ADEs and drug indications discussed by different demographic groups. For future work, we intend to further explore the application of the proposed framework to medical informatics, and particularly drug analyses, extending the work for social media–based ADE discovery discussed by different demographics groups. We would enhance the proposed framework to tackle the problem of semisupervised learning with multiple labels, rather than only a single label. We also plan to make a sentence-based ADE discovery dataset and present it publicly and make it freely available to the research community.

Acknowledgments

The authors gratefully acknowledge the support of the Center for Predictive Computational Phenotyping, supported by the National Institute of Health Big Data to Knowledge Initiative, US4AI117924; the University of Wisconsin Institute for Clinical and Translational Research (UL1TR000427) and the support of NIGMS grant 2RO1 GM097618. The authors of the paper wish to thank Richard Berg, Ryan Frahm, and Joseph Ellefson at Marshfield Clinic Research Institute and Mehdi Assefi and Sanaz Haghani at University of Georgia for their valuable contributions in big data analytics and scientific visualizations presented in this work. Special thanks goes to Anne Nikolai at Marshfield Clinic Research Institute for her valuable contributions in manuscript preparation.

Authors' Contributions

PP, DP, and APT conceived and designed the study. PP and DP secured the funding for this contribution. APT, JB, EL, and ES designed and developed the software framework. APT, JB, EL, ES, AM, JM, and ZY conducted the data collection. All authors contributed to the analysis and interpretation of the results and scientific visualization. PP and APT led the writing of this manuscript with all coauthors comments. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Selecting the relevant documents, Building the training set, Word2vec neural network model, Well-known ADEs.

[PDF File (Adobe PDF File), 640KB-Multimedia Appendix 1]

Multimedia Appendix 2

bigNN experimental validations.

[PDF File (Adobe PDF File), 195KB-Multimedia Appendix 2]

References

1. Gottlieb A, Hoehndorf R, Dumontier M, Altman RB. Ranking adverse drug reactions with crowdsourcing. J Med Internet Res 2015;17(3):e80 [FREE Full text] [doi: 10.2196/jmir.3962] [Medline: 25800813]
2. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. J Am Med Assoc 1995 Jul 05;274(1):29-34. [Medline: 7791255]
3. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. J Am Med Assoc 1997;277(4):301-306. [Medline: 9002492]
4. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. J Am Med Assoc 1998 Apr 15;279(15):1200-1205. [Medline: 9555760]
5. Trame MN, Biliouris K, Lesko LJ, Mettetal JT. Systems pharmacology to predict drug safety in drug development. Eur J Pharm Sci 2016 Oct 30;94:93-95. [doi: 10.1016/j.ejps.2016.05.027] [Medline: 27251780]
6. Liu J, Zhao S, Zhang X. An ensemble method for extracting adverse drug events from social media. Artif Intell Med 2016 Jun;70:62-76. [doi: 10.1016/j.artmed.2016.05.004] [Medline: 27431037]
7. Price J. What can big data offer the pharmacovigilance of orphan drugs? Clin Ther 2016 Dec;38(12):2533-2545. [doi: 10.1016/j.clinthera.2016.11.009] [Medline: 27914633]
8. van Walsem A, Pandhi S, Nixon RM, Guyot P, Karabis A, Moore RA. Relative benefit-risk comparing diclofenac to other traditional non-steroidal anti-inflammatory drugs and cyclooxygenase-2 inhibitors in patients with osteoarthritis or rheumatoid arthritis: a network meta-analysis. Arthritis Res Ther 2015 Mar 19;17:66 [FREE Full text] [doi: 10.1186/s13075-015-0554-0] [Medline: 25879879]

9. Pignatti F, Ashby D, Brass EP, Eichler H, Frey P, Hillege HL, et al. Structured frameworks to increase the transparency of the assessment of benefits and risks of medicines: current status and possible future directions. Clin Pharmacol Ther 2015 Nov;98(5):522-533. [doi: 10.1002/cpt.203] [Medline: 26261064]
10. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug Saf 2014 Oct;37(10):777-790 [FREE Full text] [doi: 10.1007/s40264-014-0218-z] [Medline: 25151493]
11. Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. Trials 2012 Aug 20;13:138 [FREE Full text] [doi: 10.1186/1745-6215-13-138] [Medline: 22906139]
12. Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch DW. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ISPOR risk-benefit management working group. Value Health 2010 Aug;13(5):657-666 [FREE Full text] [doi: 10.1111/j.1524-4733.2010.00725.x] [Medline: 20412543]
13. Eichler H, Abadie E, Raine JM, Salmonson T. Safe drugs and the cost of good intentions. N Engl J Med 2009 Apr 02;360(14):1378-1380. [doi: 10.1056/NEJMp0900092] [Medline: 19339718]
14. Garrison LP, Towse A, Bresnahan BW. Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. Health Aff (Millwood) 2007;26(3):684-695 [FREE Full text] [doi: 10.1377/hlthaff.26.3.684] [Medline: 17485745]
15. Jonnagaddala J, Jue T, Dai H. Binary classification of Twitter posts for adverse drug reactions. 2016 Jan Presented at: In Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing; 2016; Big Island, HI. USA p. 4-8.
16. Rastegar-Mojarad M, Elayavilli R, Wang L, Prasad R, Liu H. Prioritizing Adverse Drug Reaction and Drug Repositioning Candidates Generated by Literature-Based Discovery. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2016 Presented at: BCB '16; October 02 - 05, 2016; Seattle, WA, USA p. 289-296. [doi: 10.1145/2975167.2975197]
17. Yang M, Kiang M, Shang W. Filtering big data from social media--building an early warning system for adverse drug reactions. J Biomed Inform 2015 Apr;54:230-240 [FREE Full text] [doi: 10.1016/j.jbi.2015.01.011] [Medline: 25688695]
18. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015 Feb;53:196-207 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.002] [Medline: 25451103]
19. Liu X, Chen H. AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. 2013 Presented at: International Conference on Smart Health; August 3, 2013; Beijing, China p. 134-150.
20. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the 2012 international workshop on Smart health and wellbeing. 2012 Oct 29 Presented at: SHB '12; October 29 - 29, 2012; Maui, Hawaii, USA p. 25-32 URL: http://europepmc.org/abstract/MED/28967001 [doi: 10.1145/2389707.2389713]
21. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. AMIA Annu Symp Proc 2011;2011:217-226 [FREE Full text] [Medline: 22195073]
22. Paul M, Dredze M. Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models. 2013 Presented at: Proceedings of NAACL-HLT 2013; 9-14 June, 2013; Atlanta, Georgia p. 168-178.
23. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med 2012 Mar 14;4(125):125ra31 [FREE Full text] [doi: 10.1126/scitranslmed.3003377] [Medline: 22422992]
24. Liang Z, Huang JX, Zeng X, Zhang G. DL-ADR: a novel deep learning model for classifying genomic variants into adverse drug reactions. BMC Med Genomics 2016 Aug 10;9(Suppl 2):48 [FREE Full text] [doi: 10.1186/s12920-016-0207-4] [Medline: 27510822]
25. Iyer SV, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. J Am Med Inform Assoc 2014;21(2):353-362 [FREE Full text] [doi: 10.1136/amiajnl-2013-001612] [Medline: 24158091]
26. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 2012 Jun;91(6):1010-1021 [FREE Full text] [doi: 10.1038/clpt.2012.50] [Medline: 22549283]
27. Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics 2012;3(1):15 [FREE Full text] [doi: 10.1186/2041-1480-3-15] [Medline: 23256479]
28. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010;160(Pt 1):739-743. [Medline: 20841784]
29. NCBI. PubMed URL: https://www.ncbi.nlm.nih.gov/pubmed [accessed 2017-11-29] [WebCite Cache ID 6vLTyklNq]
30. Naples JG, Hanlon JT, Schmader KE, Semla TP. Recent literature on medication errors and adverse drug events in older adults. J Am Geriatr Soc 2016 Feb;64(2):401-408 [FREE Full text] [doi: 10.1111/jgs.13922] [Medline: 26804210]
31. Noguchi C, Sakuma M, Ohta Y, Bates DW, Morimoto T. Prevention of medication errors in hospitalized patients: the Japan adverse drug events study. Drug Saf 2016 Nov;39(11):1129-1137. [doi: 10.1007/s40264-016-0458-1] [Medline: 27638660]
32. Meier F, Maas R, Sonst A, Patapovas A, Müller F, Plank-Kiegele B, et al. Adverse drug events in patients admitted to an emergency department: an analysis of direct costs. Pharmacoepidemiol Drug Saf 2015 Feb;24(2):176-186. [doi: 10.1002/pds.3663] [Medline: 24934134]

33. Nuckols TK, Smith-Spangler C, Morton SC, Asch SM, Patel VM, Anderson LJ, et al. The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. Syst Rev 2014 Jun 04;3:56 [FREE Full text] [doi: 10.1186/2046-4053-3-56] [Medline: 24894078]
34. Kanaan AO, Donovan JL, Duchin NP, Field TS, Tjia J, Cutrona SL, et al. Adverse drug events after hospital discharge in older adults: types, severity, and involvement of Beers Criteria Medications. J Am Geriatr Soc 2013 Nov;61(11):1894-1899 [FREE Full text] [doi: 10.1111/jgs.12504] [Medline: 24116689]
35. Dormann H, Sonst A, Müller F, Vogler R, Patapovas A, Pfistermeister B, et al. Adverse drug events in older patients admitted as an emergency: the role of potentially inappropriate medication in elderly people (PRISCUS). Dtsch Arztebl Int 2013 Mar;110(13):213-219 [FREE Full text] [doi: 10.3238/arztebl.2013.0213] [Medline: 23596501]
36. Finkelstein J, Chen Q, Adams H, Friedman C. Automated summarization of publications associated with adverse drug reactions from PubMed. AMIA Jt Summits Transl Sci Proc 2016;2016:68-77 [FREE Full text] [Medline: 27570654]
37. Natarajan S, Bangera V, Khot T, Picado J, Wazalwar A, Costa VS, et al. Markov logic networks for adverse drug event extraction from text. Knowl Inf Syst 2017 May;51(2):435-457. [doi: 10.1007/s10115-016-0980-6] [Medline: 29123330]
38. Kang N, Singh B, Bui C, Afzal Z, van Mulligen EM, Kors JA. Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics 2014 Mar 04;15:64 [FREE Full text] [doi: 10.1186/1471-2105-15-64] [Medline: 24593054]
39. NCBI. PubMed Central URL: https://www.ncbi.nlm.nih.gov/pmc/ [accessed 2017-11-29] [WebCite Cache ID 6vLUJCock]
40. MedHelp. URL: http://medhelp.org/ [accessed 2017-11-29] [WebCite Cache ID 6vLUQL3F0]
41. Patient. URL: https://patient.info/ [accessed 2017-11-29] [WebCite Cache ID 6vLUVsMfF]
42. WebMD. URL: https://www.webmd.com/default.htm [accessed 2017-11-29] [WebCite Cache ID 6vLUcTEob]
43. Scrapy. URL: https://scrapy.org/ [accessed 2017-11-29] [WebCite Cache ID 6vLUjwo7C]
44. Elastic. Elasticsearch URL: https://www.elastic.co/ [accessed 2017-11-29] [WebCite Cache ID 6vLUpqeIv]
45. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J Biomed Inform 2015 Aug 17;57:333-349 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.013] [Medline: 26291578]
46. Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short text classification in twitter to improve information filtering. 2010 Presented at: The 33rd international ACM SIGIR conference on Research and development in information retrieval; Jul 19, 2010; Geneva, Switzerland p. 841-842.
47. Sivic J, Zisserman A. Efficient visual search of videos cast as text retrieval. IEEE Trans Pattern Anal Mach Intell 2009 Apr;31(4):591-606. [doi: 10.1109/TPAMI.2008.111] [Medline: 19229077]
48. Sahlgren M, Cöster R. Using bag-of-concepts to improve the performance of support vector machines in text categorization. 2004 Presented at: the 20th International Conference on Computational Linguistics; August 23, 2004; Geneva, Switzerland.
49. Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C. Text classification using string kernels. J Mach Learn Res 2002;2:419-444.
50. Pedersen T. A decision tree of bigrams is an accurate predictor of word sense. 2001 Presented at: NAACL '01 Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies; June 01 - 07, 2001; Pittsburgh, Pennsylvania p. 1-8.
51. Scott S, Matwin S. Feature engineering for text classification. 1999 Presented at: ICML '99 Proceedings of the Sixteenth International Conference on Machine Learning; June 27 - 30, 1999; San Francisco, CA, USA p. 379-388.
52. Mikolov T, Chen K, Corrado G, Dean J. arxiv.org. 2013. Efficient estimation of word representations in vector space URL: https://arxiv.org/pdf/1301.3781.pdf [accessed 2017-11-30] [WebCite Cache ID 6vNQxsjjp]
53. Bansal M, Gimpel K, Livescu K. Tailoring Continuous Word Representations for Dependency Parsing. In ACL (2) 2014 Jun (pp. 809-815) 2014.
54. Dyer C, Ballesteros M, Ling W, Matthews A, Smith N. Transition-based dependency parsing with stack long short-term memory. arXiv preprint arXiv.08075 May 2015:29.
55. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Statnlp.org. 2016. Neural architectures for named entity recognition URL: http://www.statnlp.org/wp-content/uploads/2017/06/nner.pdf [accessed 2017-11-30] [WebCite Cache ID 6vNRB665R]
56. Santos CN, Guimaraes V. Aclweb.org. 2015. Boosting named entity recognition with neural character embeddings URL: http://www.anthology.aclweb.org/W/W15/W15-3904.pdf [accessed 2017-11-30] [WebCite Cache ID 6vNRHxjLX]
57. Lilleberg J, Zhu Y, Zhang Y. Support vector machines and word2vec for text classification with semantic features. 2015 Presented at: Cognitive Informatics & Cognitive Computing (ICCI*CC), 2015 IEEE 14th International Conference on; July 6, 2015; Beijing, China p. 136-140.
58. Zhang X, Zhao J, LeCun Y. Courant Institute of Mathematical Sciences, New York University. 2015. Character-level convolutional networks for text classification URL: https://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf [accessed 2017-11-30] [WebCite Cache ID 6vNRQDFcC]
59. Xue B, Fu C, Shaobin Z. A new clustering model based on Word2vec mining on Sina Weibo users' tags. IJGDC 2014;7(3):41-48. [doi: 10.14257/ijgdc.2014.7.3.05]

60. Goldberg Y, Levy O. Arxiv.org. 2014. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method URL: https://arxiv.org/pdf/1402.3722.pdf [accessed 2017-12-01] [WebCite Cache ID 6vNReVCAM]
61. Rong X. arxiv.org. 2014. word2vec parameter learning explained URL: https://arxiv.org/pdf/1411.2738.pdf [accessed 2017-12-01] [WebCite Cache ID 6vNRhUI42]
62. Ye J. Cosine similarity measures for intuitionistic fuzzy sets and their applications. Math Comput Model 2011;53(1-2):91-97. [doi: 10.1016/j.mcm.2010.07.022]
63. Muflikhah L, Baharudin B. Document clustering using concept space and cosine similarity measurement. 2009 Presented at: Computer Technology and Development, 2009. ICCTD '09. International Conference on; November 13, 2009; Kota Kinabalu, Malaysia p. 58-62.
64. cTAKES: clinical Text Analysis and Knowledge Extraction System. URL: http://ctakes.apache.org/ [accessed 2017-11-29] [WebCite Cache ID 6vLTlneFL]
65. Apache Spark. URL: https://spark.apache.org [accessed 2017-11-29] [WebCite Cache ID 6vLVIS1gd]
66. The University of Waikato. Weka Library URL: https://www.cs.waikato.ac.nz/ml/weka/ [accessed 2017-11-29] [WebCite Cache ID 6vLUwaBCP]
67. Hadoop. URL: http://hadoop.apache.org/ [accessed 2017-11-29] [WebCite Cache ID 6vLVBwzjO]

Abbreviations

ADE: adverse drug event
AUC: area under the curve
BigNN: big data neutral network
BoW: bag-of-words
EHR: electronic health record
EP: epoch
ITR: iteration
MWF: minimum word frequency
NLP: natural language processing
POS: part-of-speech
SVM: support vector machine
WS: window size
XML: extensible markup language

Edited by G Eysenbach; submitted 13.10.17; peer-reviewed by V Hristidis, R Segall, F Rohani; comments to author 03.11.17; revised version received 07.11.17; accepted 08.11.17; published 08.12.17

Please cite as:
P Tafti A, Badger J, LaRose E, Shirzadi E, Mahnke A, Mayer J, Ye Z, Page D, Peissig P
Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure
JMIR Med Inform 2017;5(4):e51
URL: http://medinform.jmir.org/2017/4/e51/
doi: 10.2196/medinform.9170
PMID: 29222076