An efficient chemical drug contextual similarity based classification model on Biomedical document datasets

[1]K.S.S.Joseph Sastry, [2]M.Sree Devi, P, [3]K.Raja sekhar, P

Department of CSE, Research Scholar, K.L.E.F, Guntur
rofessor, Department of CSE, Amritha sai institute of sciences
rofessor ,Department of CSE , K.L.E.F ,Guntur

Abstract

The exponential growth of biological literature has led to the accumulation of vast knowledge, encompassing various areas such as protein-protein interactions (PPIs), chemical-to-drug interactions, and drug-to-drug interactions (DDIs). Automatically identifying and categorizing biomedical associations provides significant advantages in diverse biomedical research fields. Over the past decade, notable progress has been made in identifying biomedical relationships. However, traditional models have primarily focused on PPIs and DDIs, often neglecting chemical-to-drug relations due to challenges related to extensive training data and accurate classification models.This study introduces a novel approach for extracting different chemical-to-drug relationships using a hybrid filtered-based text classification model. The proposed model incorporates a new measure of similarity between chemicals and drugs, as well as a maximized kernel learning-based SVM technique, which aims to identify crucial patterns within the training data. Experimental results demonstrate that the presented chemical-to-drug classification model outperforms existing interaction models in terms of efficiency.

Keywords: chemical names, drug names, biomedical, support vector machine.

1.Introduction

Ensemble-based approaches have gained prominence in recent bioinformatics research. These methods have found applications in various areas, such as classifying gene-gene-gene interaction microarray data and predicting gene-gene function. In traditional learning machine problems, a single algorithm is developed and compared against others. Meta-learning, an ensemble-learning technique, allows the system to learn the overall accuracy of predictions from multiple trained algorithms [1]. This approach leverages the knowledge gained from individual algorithms, referred to as meta-knowledge, to enable successful predictions. Meta-learning draws inspiration from human decision-making, incorporating diverse perspectives and viewpoints.Drug interactions occur when two or more drugs are administered together, affecting each other's effects and functions in the human body. Such interactions can lead to unintended side effects or alter the therapeutic effects of medications. For example, aspirin, when taken alongside blood thinners like Warfarin (Coumadin), can increase the risk of excessive bleeding. With the increasing number of approved drugs in recent years, the prevalence of individuals with multiple chronic conditions has also risen. Approximately 71% of total medical expenditure in the United States is related to the healthcare of individuals with multiple chronic conditions [2]. Consequently, polypharmacy, which involves the use of four or more medications by an

individual, has become increasingly common, particularly among adults over the age of 65. The number of prescribed medications continues to grow as medical knowledge evolves and new discoveries are made. However, this valuable information is often scattered across various sources, including journal articles and patient records. Statistical methods can be employed to collect and integrate data from these disparate text sources, allowing for a comprehensive understanding of different disease features (e.g., medications or symptoms) and their changes over time [1-3]. Up-to-date disease profiles can be beneficial for various applications, such as decision support, quality control, and answering patient queries.In the context of differentiating between curative and incurable terms, a support vector machine (SVM), which is a powerful statistical learning theory classification algorithm, is employed. SVMs are widely used to solve supervised learning classification problems, which involve training samples with multiple features and assigned class labels. By defining an SVM model, prototypes are created and the features of unknown instances are examined to accurately determine their category labels. The underlying principle of SVM is to identify a description that effectively separates the two test groups. As practice samples can have numerous features, resulting in a high-dimensional feature space, the hyperplane generated by SVM distinctly differentiates information into separate classes.Named entity recognition (NER) not only serves the purpose of retrieving information but also provides information at a higher level. In the NER action term of CoINNER, every action term represents a primary subject learned by the CTD course from various articles (subtopics). To construct keywords for action verbs, Latent Dirichlet Allocation (LDA), a generative statistical model, is employed. LDA enables the extraction of keywords for subtopics once the LDA model has been computed. For the NER action term in CoINNER, the top ten keywords for each subject are selected, with overlapping keywords among subjects eliminated. Ultimately, the action terms are denoted by mapping keywords to the respective subjects.For the triage task in the BioCreative IV CTD Track, participants were provided with the learning dataset, which had been curated by CTD biocurators and comprised 1,112 papers [4]. The dataset, available in XML-based files, contained relevant curated information such as PubMed IDs, names, abstracts, gene, chemical, disease, and action word annotations [4]. CTD staff manually supervised a sample dataset consisting of 510 records. The research dataset included 1,122 genes, 1,192 chemical compounds, 943 disorders, 966 chemical/gene-specific conditions, and 3,953 manually curated interactions.Clinical drug interactions (DDIs) pose a significant risk, leading to morbidity and mortality, and require accurate prediction and identification. Various stakeholders, including the pharmaceutical industry, drug regulatory agencies, healthcare professionals, and patients, are particularly concerned about DDI prediction. DDIs occur when one medication alters the pharmacokinetics or pharmacodynamic properties of another drug. Most DDI predictions rely on careful evaluation of

molecular targets and enzyme metabolism, often focusing on a limited number of drug pairs or specific metabolizing enzymes, such as P450 enzymes. However, these techniques are costly and limited in scope. Some studies have conducted domain-specific DDI investigations, but generalizing the prediction of future DDIs across different contexts remains challenging. Recent research efforts have leveraged data collection and statistical methods to improve DDI identification [5].

2. Related works

Meta-learning encompasses two central approaches: generalization and generalisation [6]. In stacked generalization, base learners are trained on the same dataset, and their outputs are combined into a meta-model at a higher level. Cascade generalization involves using the original set of functions for the output of the basic learner and passing it to the subsequent basic learner in a sequential manner. Predicting drug-target interactions relies on obtaining necessary information from validated experimental data, literature search, and computer modeling. The Medline Database can be mined using optimized algorithms to investigate various drug interaction problems [7]. Another commonly employed technique is in silico modeling, which involves simulating compound-protein interactions through docking simulations [8]. High-performance computers can be utilized to accurately predict the communication network of drug candidates and classify clinically significant adverse drug reactions (ADRs). However, molecular docking methods suffer from long calculation times and limited accuracy. Threading/structural FINDSITE-based methods have been developed to scan the entire proteome against various compounds, offering better accuracy in identifying drug-target interactions than traditional docking methods.The model developed in this study incorporates features from the aforementioned models to define and explain important aspects of drug-drug interactions (DDIs), such as the effect of the interaction and factors that increase or reduce DDI risks. However, it does not provide a representative mechanism through which the communication occurs. To extract drug-related knowledge from texts and populate an ontology, an ontology-based template was implemented. While previous models aimed to represent DDIs in a general sense, they did not adequately capture the diverse mechanisms underlying DDIs. DDIs are typically classified into two main groups based on their pathways: pharmacokinetic (PK) and pharmacodynamic (PD) interactions [9]. PK interactions occur when a drug affects the concentration of another drug, while PD interactions occur when a drug modifies the effects of another drug without changing its

concentration, often by acting on the same target. The following models focus on describing the mechanisms of PK and PD DDIs [10].

Significant interactions between drugs in elderly Indian patients have been confirmed, and preventable interactions with medicines contribute to 3 to 5% of all medical errors in hospitals. Some high-profile drugs have been withdrawn from the market due to DDIs, highlighting the importance of evaluating drugs for potential interactions. Support Vector Regression (SVR) is a powerful and efficient learning algorithm that maps non-linear functions into a high-dimensional space through non-linear projection. SVR selects the appropriate kernel function without requiring knowledge of the specific non-linear mapping type. The selection of kernel functions in SVR theory is an actively researched area, with cross-validation approaches and modification of the SVR kernel function based on feedback weight values being proposed [11]. A cluster-based SVR with specific kernel methods has also been introduced, incorporating the Optimal Relaxation Factor (ORF) function from speech recognition. Additionally, the conventional QSAR model considers suitable reliability and incorporates details such as probability distribution [12].

3.Proposed Model

Data and Filtering

Concepts Theory refers to a scientific concept or expression. Concepts Biomedical are usually limited according to the data source underlying. Take a subject called "Parkinson's Disease," for instance, which can be applied to as definition as genes, proteins, pharmaceuticals, symptoms and other related diseases. In other words, definition can be considered the set of meaningful terms describing the object of a text for the purposes of LBD. A text article contains many ways to extract concepts. Several studies [13] want to extract biomedically relevant concepts from free text by using structured vocabulary, such as the UMLS. Additionally, some people choose to use Clinical Subject Headings (MeSH) to reflect information. MeSH words belong to the standardized vocabulary of the National Library of Medicine (NLM), which human professionals use to index citations. Specific LBD structures use MeSH words to discover information are[14].Semantic similarity of concept level can usually be calculated by methods based on knowledge and on corpus. Knowledgement-based approaches are mostly employed to show their semantine similarity between concepts of information resources[15]. Instead of using the graph distance directly in the information sources, a few generated concept vectors for conceptual similitude according to a set of ontological properties. Methods based on corpus presume that words of similar importance frequently occur in similar contexts, Latent Semantic Analysis (LSA) (16) represents words in corpus matrix as compact vectors through a single value decomposition (SVD), and GloVe reduced computation costs by direct training on non-zero elements of the corpus matrix. The common character p-grams between two strings are based on all string kernels used here. Using string kernels, the corresponding form of learning is completely separate from one language since the text is interpreted as a symbol sequence (string). Theory and analytical theories restrict the space of the methods operating at or above word scales. For example, only functions that reflect different types of assessment groups or just certain terms, such as opinion-oriented words, can be chosen. Such features fit very well for some activities, but there may be other good features not considered. The string nodes simply insert texts into an extremely wide space, provided by the substrates of length p. by assigning different weights to such functions, leaving to the learning algorithm the choice of important characteristics for the specific job. Because the space for features is not limited by any language theory, the approach to string kernel is linguistic neutral. In addition, no aspects of natural language such as verbs, phrases or context are specifically included in the process,

contrary to the standard NLP approach. An interesting observation is, however, that such features can be contained in the p-grams derived implicitly. In other words, it may prove very difficult to extract these features specifically for less well understood or low-resource languages and language-dependent methods that rely on linguistic features. Even a system that treats words as characteristic cannot be entirely independent of language, as the meaning of a word must be linguistic. A way that uses only opinion-driven words as a function is also not an independent language because it requires a collection of language-specific opinion-driven words. The approach relies on a partial speech tagger that may not be usable for certain languages, when features, such as partial speech tags, are used. On the contrary, a number of positive and negative examples are only needed for our process, so that you can acquire them easily automatically in product or movie opinions, possibly for any language studied online. Instances involving drug pairs manually annotated as interacting (either in positive or negative modalities), have been classified as positive examples of extraction; all other pairs have been marked as negative instances.

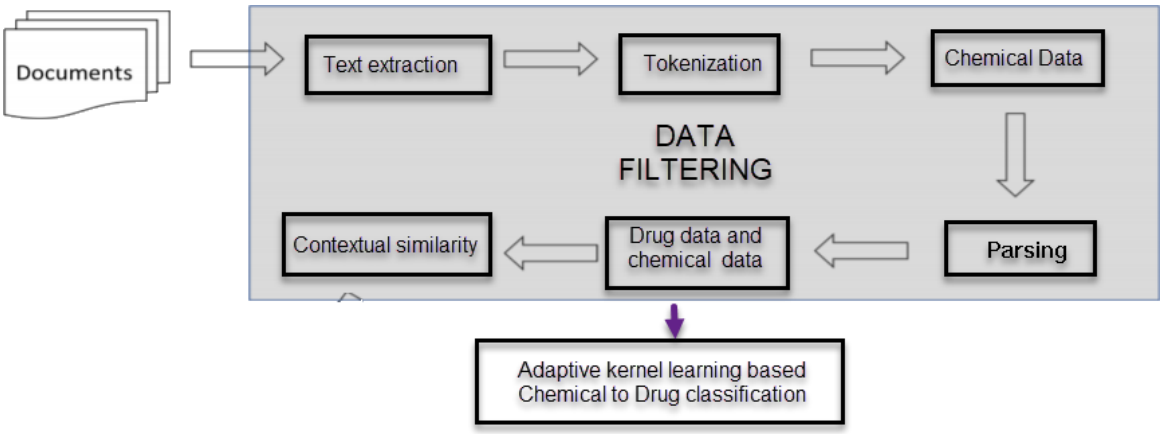

Figure 2 Proposed Model

Every instance was pre-processed before indexing features for machine learning. The numbers (e.g. "1," "34," "5.2," etc.) are replaced by "digit" to maximize their importance for a case-by-case learning algorithm. Instead of studying unique numerical phrases (for example the term "num mg" may be significant, "10 mg" may not be significant). This has helped the algorithm to equate numerical comparisons with each other in a general pattern. Similarly, the names of pharmaceutical drugs, active ingredients and metabolites have been substituted with the string "drug name" to explain those names in each sentence. It requires a learning

algorithm to generalize interaction participants so that they cannot distinguish experiences based on the identity of the participants. Another issue was that multiple mentions of the same substance within a sentence of an example, such as the sentence ' A and B management contributes to increased AUC rates for B' To order to mitigate these problems, two solutions are introduced. The dataset was first compressed into a sequence of cases where single pairs of drugs and sentences were used. In the case of one of the reference cases that contributed to the condensed example, the condensed instance represented an interaction. In such a way, multiple drug statements within a phrase containing an interaction will result in the interaction between the two drugs representing one example. Second, the Stanford NLP Parser

, and the Clear Parser have been used to extract extra functions for each example from the sentencing document as shown in fig 2.

The contextual similarity between chemical and drug is given by

$$\mathrm{Sim}_{CD}(S_1, S_2) = Max\{hybridMI(S_1, S_2), \frac{V_1 \cdot V_2}{\|V_1\| \cdot \|V_2\|}\}$$

$$hybridMI(\mathrm{C}_1; \mathrm{D}_2) = \sum_{w} P(w' \mid \mathrm{C}) \log \frac{P(w' \mid \mathrm{C})}{P(w' \mid \mathrm{D})}$$

Advanced Kernel learning based Bayesian text classification

Each text d is interpreted as a vector x in the space of the n-dimensional characteristic vector when classifying text according to functional collection of pre-treated samples (such as word segmentation punctuation). There are currently several text classification machine learning algorithms such as the k-closest approach and SVM algorithm etc. The SVM algorithm reflects the current classification method's performance level. However, gradual changes to the SVM classification models not only can add new samples or information categories to the classification system but do not impact the classifier's quality and maintain number of categories. The feature objects are usually words or sentences. Words as the attribute are usually better than characters and phrases according to the experiments as shown in fig 3. Therefore, if the text is represented in the vector field as a vector, the word segmentation is performed first in the text and segmented words are used as a vector element. In other words, if the word appears in the message, the text vector dimension is 0. This approach, however, does not reflect the degree of word position in the text and thus 01 is replaced slowly by a more specific word rate. The frequency of word and relativity is separated into absolute frequency.

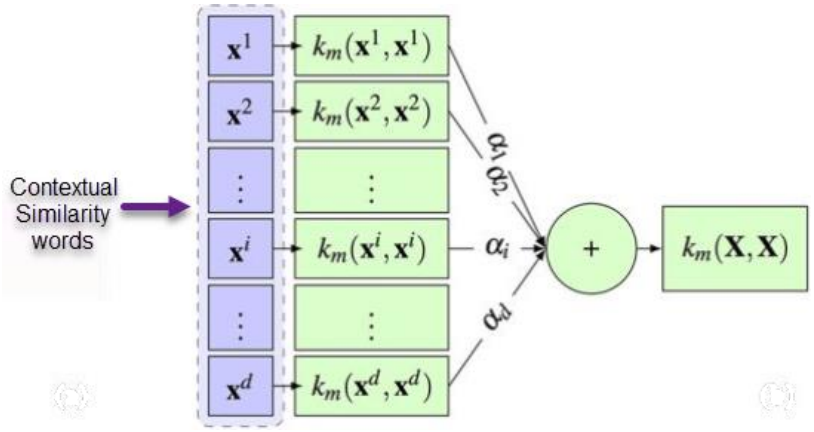

Figure 3: Kernel mapping for SVM

Advanced integrated maximized kernel function used in the SVM model is given below

$$K_1(\mathrm{p},\mathrm{q}) = \tanh(\mathrm{x}\ p \cdot q + y)$$

$$K_2(\mathrm{p},\ \mathrm{q}) = (1 + p \cdot q)^d$$

$$K_3(\mathrm{p},\mathrm{q}) = x[\exp(\frac{1}{\|\mathrm{p} - q\|^2 + \square^2}) - 1]$$

$$\mathrm{MaxAcc} = \{K_1(\mathrm{p},\ \mathrm{q}), K_2(\mathrm{p},\ \mathrm{q}), K_3(\mathrm{p},\ \mathrm{q})\}$$

In this study, we have assessed methods of selection based on the weight vector of the SVM.Since, these methods of feature selection use data on which features contribute most to the category, they are supposed to significantly reduce features. Intuitively, a method of selecting features based on SVM before SVM is a desirable solution because the task and classification used the same model of decision.

4. Experimental results

The probability matrix for training data consists of chemical symbols and drug names. We developed the classifier to predict the new associations of drug phenotypes. We calculated classification accuracy and three co-occurrence-based methods in order to evaluate the classifier's efficiency. The three forms are (1) co-occurrence number phrases, (2) co-occurrence binary phrases and (3) co-occurrence numerical frequency frequency-inverse report word (TF IDF) expression. Co-occurrence is the most simple way of mining text, and it is believed to be linked functionally when two words simultaneously appear in the same text. Method (1) tests the co-occurrence rate of drugs and genes in a sentence; method (2) decides whether drugs and genes have been detected in one sentence or not. Form (3) is the result of normalization by TF-IDF of the form matrix (1). TF-IDF is a word frequency item and inverse document frequency that determines the meaning of a term in a specific document in the category of documents. The mean squared error (MSE), maximized absolute

error (MAE) and the squared correlation coefficient between the sample data and the test results are considered in order to make a comparison between the output of the models.

$$MAE = \max(|t_t - t_p|)$$

$$MSE = \frac{1}{n}\sum_{t}(t - t_p)^2$$

$$R^2 = \frac{n\sum t_t^T t_p - \sum t_t \sum t_p}{\sqrt{n\sum t_t^T t_t - (\sum t_t)^2}\sqrt{n\sum t_p^T t_p - (\sum t_p)^2}}$$

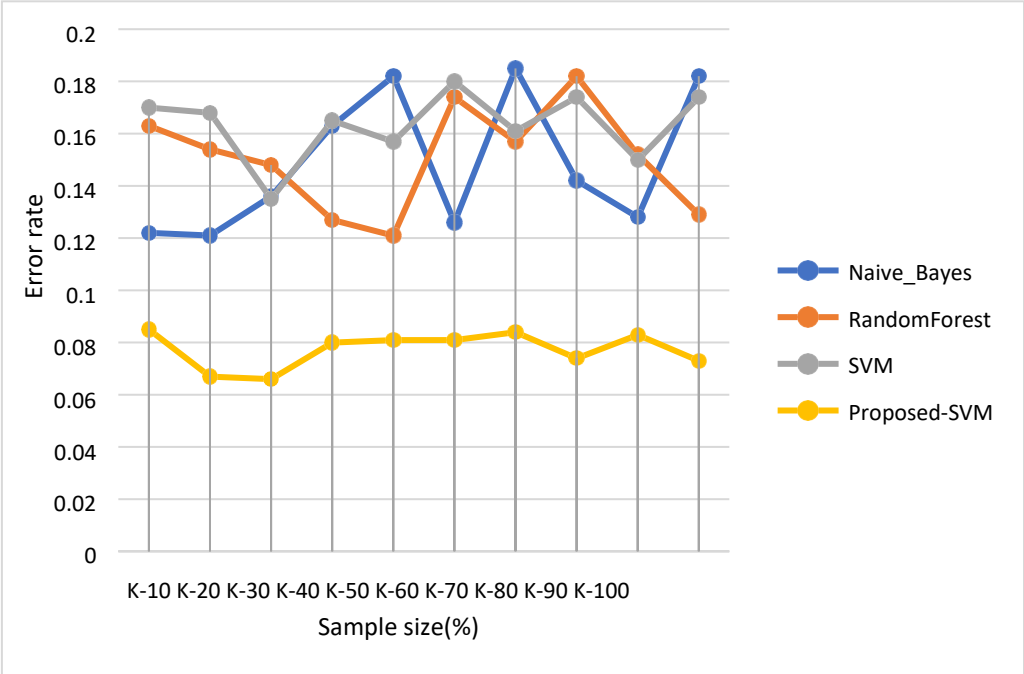

Figure 4: Performance analysis of proposed SVM model to traditional text mining models for error rate

Table 1: Performance analysis of proposed SVM model to traditional text mining models for accuracy

SampleSize(%)	Naive_Bayes	RandomForest	SVM	Proposed-SVM
D-10	0.836	0.861	0.911	0.967
D-20	0.834	0.862	0.897	0.959
D-30	0.838	0.875	0.921	0.962
D-40	0.846	0.859	0.922	0.958
D-50	0.847	0.88	0.912	0.962
D-60	0.84	0.861	0.909	0.963
D-70	0.843	0.879	0.916	0.961
D-80	0.834	0.863	0.924	0.966
D-90	0.834	0.878	0.897	0.958
D-100	0.841	0.873	0.901	0.961
SampleSize(%)	Naive_Bayes	RandomForest	SVM	Proposed-SVM
#10	8301.98	8026.08	6966	6572.84
#20	8633.28	8097.94	6992.27	5917.93
#30	8183.6	8180.58	7158.48	5968.33
#40	8599.71	8098.55	7102.89	6440.74
#50	8176.76	7987.1	7161.02	6238.45
#60	8775.64	7899.71	6918.17	6084.15
#70	8867.98	8152.43	7562.74	6368.19
#80	8788.26	7974.35	7214.1	6328.78
#90	7996.45	7983.75	7499	5892.23
#100	8575.77	8144.6	7589.78	6244.45

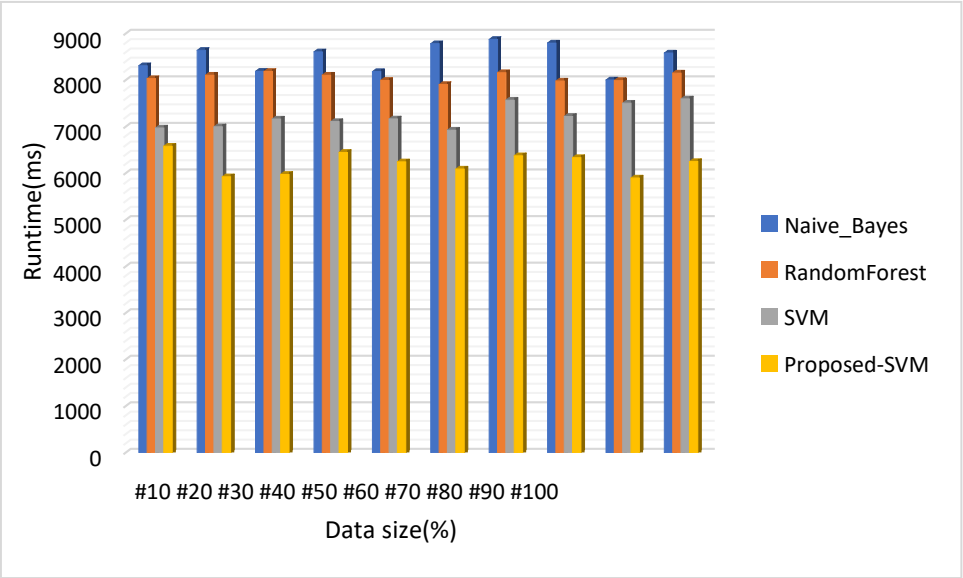

Figure 5: Performance analysis of proposed SVM model to traditional text mining models for runtime(ms).

Conclusion

Many approaches to the biomedicine of text mining do not deal with specific patterns of cause-effect that justify the results. In this report, an efficient new model for text mining from biomedical literature helps to identify theories of causal effects associated with diseases, medicines, etc. While electronic health records are becoming increasingly prevalent, a great deal of data about patients ' health is still documented within unstructured documents. The focus of natural language processing (NLP) research has been on the interpretation of these texts for a number of years with impressive results but further work needs to be carried out. Knowledge of medicine is important not only in understanding patient health (for example, interactions with pharmaceutical products or interactions with drug-enzymes), but also in secondary applications, such as treatment efficacy testing.In this work, different chemical to drug relationships are extracted using the hybrid filtered based text classification model. In this model, a new chemical to drug similarity measure and maximized kernel learning based SVM technique is proposed to find the essential key patterns in the training data. Experimental results proved that the present chemical to drug classification model is efficient than the existing interaction models.

References

[1]R. McEntire et al., "Application of an automated natural language processing (NLP) workflow to enable federated search of external biomedical content in drug discovery and development," Drug Discovery Today, vol. 21, no. 5, pp. 826–835, May 2016.

[2]S. Ayvaz et al., "Toward a complete dataset of drug–drug interaction information from publicly available sources," Journal of Biomedical Informatics, vol. 55, pp. 206–217, Jun. 2015.

[3]M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, and T. Declerck, "The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions," Journal of Biomedical Informatics, vol. 46, no. 5, pp. 914–920, Oct. 2013.

[4]T. Ly et al., "Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology," Journal of Biomedical Informatics, vol. 83, pp. 73–86, Jul. 2018.

[5]K. Negi, A. Pavuri, L. Patel, and C. Jain, "A novel method for drug-adverse event extraction using machine learning," Informatics in Medicine Unlocked, p. 100190, May 2019.

[6]L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal, "Natural language processing for aviation safety reports: From classification to interactive analysis," Computers in Industry, vol. 78, pp. 80–95, May 2016.

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

[8]R. McEntire et al., "Application of an automated natural language processing (NLP) workflow to enable federated search of external biomedical content in drug discovery and development," Drug Discovery Today, vol. 21, no. 5, pp. 826–835, May 2016.

[9]Y. Zhang et al., "Neural network-based approaches for biomedical relation classification: A review," Journal of Biomedical Informatics, vol. 99, p. 103294, Nov. 2019.

[10]Z. Khalid and O. U. Sezerman, "ZK DrugResist 2.0: A TextMiner to extract semantic relations of drug resistance from PubMed," Journal of Biomedical Informatics, vol. 69, pp. 93–98, May 2017.

[11]A. Coden, D. Gruhl, N. Lewis, M. Tanenblatt, and J. Terdiman, "SPOT the Drug! An Unsupervised Pattern Matching Method to Extract Drug Names from Very Large Clinical Corpora," in 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, 2012, pp. 33–39.

[12]S. Santiso, A. Pérez, and A. Casillas, "Exploring Joint AB-LSTM With Embedded Lemmas for Adverse Drug Reaction Discovery," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 5, pp. 2148–2155, Sep. 2019.

[13]Z. Zeng, Y. Deng, X. Li, T. Naumann, and Y. Luo, "Natural Language Processing for EHR-Based Computational Phenotyping," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 1, pp. 139–153, Jan. 2019.

[14]H. Huang et al., "Discovering Medication Patterns for High-Complexity Drug-Using Diseases Through Electronic Medical Records," IEEE Access, vol. 7, pp. 125280–125299, 2019.

[15]B. Ru, D. Li, Y. Hu, and L. Yao, "Serendipity—A Machine-Learning Application for Mining Serendipitous Drug Usage From Social Media," IEEE Transactions on NanoBioscience, vol. 18, no. 3, pp. 324–334, Jul. 2019.

[16]J. Atkinson and A. Rivas, "Discovering Novel Causal Patterns From Biomedical Natural-Language Texts Using Bayesian Nets," IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 6, pp. 714–722, Nov. 2008.

[17].Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance .Potharaju, S.P. and Sreedevi, M.

[18].A Novel M-Cluster of Feature Selection Approach Based on Symmetrical Uncertainty for Increasing Classification Accuracy of Medical Datasets Sai Prasad Potharaju and M.Sreedevi Dept of CSE , K L University, Guntur, Andhra Pradesh, India Potharaju, S.P andSreedevi,M

[19] ARPN Journal of Engineering and Applied SciencesOpen AccessVolume 13, Issue 9, 1 May 2018, Pages 3129-3135 Prototype analysis of different data mining classification and clustering approaches Kolli, S. and Sreedevi, M.

[20] Gazi University Journal of Science Volume 31, Issue 3, 2018, Pages 775-787 Correlation coefficient based feature selection framework using graph construction (Article) Potharaju, S.P.and Sreedevi, M.

[21] Incremental mining for regular frequent patterns in vertical format, G.Vijay Kumar and Valli Kumari.V International Journal of Engineering and Technology Volume 5, Issue 2, 2013, Pages 1506-1511.

[22] Mining popular patterns from multidimensional database Vijay Kumar, G. and Krishna Chaitanya, Indian Journal of Science and Technology Volume 9, Issue 17, 1 May 2016, Article number 93106

[23] Temperature and heart beat monitoring system using IOT Vijay Kumar, G.,Bharadwaja, A. and Nikhil Sai, Proceedings - International Conference on Trends in Electronics and Informatics, ICEI 2017Volume 2018-January, 21 February 2018, Pages 692-695

[24] "A Document Ranking Approach based on Weighted-Gene/Protein in Large Biomedical Documents using MapReduce Framework" K.S.S. Joseph Sastry, Dr.Venkata Daya Sagar Ketaraju, DOI 10.5013/IJSSST.a.19.06.26 ISSN: 1473-804x , 1473-8031.

[25] "A Mutual Conditional Probability based document ranking model using Map-Reduce Framework" K.S.S. Joseph Sastry, Dr.M.Sree Devi, International Journal of Advanced Science and Technology Vol. 28, No. 15, (2019), pp. 493-500.

[26] Kolli srinivas, M.Sreedevi,"A novel index based procedure to explore similar attribute similarity in uncertain categorical data",ARPN Journal of Engineering and Applied Sciences,2019.

[27] V. Laxmi Narasamma and M. Sreedevi,"A Framework to Analysis of Tweets using Multi-Level Tree Algorithms",Journal of Advanced Research in Dynamical and Control Systems Vol. Sp– 18 ,2017.

[28] V. Laxmi Narasamma and M. Sreedevi,"A Framework to Analysis of Tweets using Multi-Level Tree Algorithms",Journal of Advanced Research in Dynamical and Control Systems Vol. Sp– 18 ,2017.