International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 20s, 2025
https://theaspd.com/index.php

Optimizing Adverse Drug Reaction Surveillance: Integrating Latent Semantic Analysis And Artificial Neural Networks In Pharmacovigilance

M. Umamaheswari[1], Dr. P. Ranjana[2]
[1,2]CSE Department, Hindustan Institute of Technology and Science
[1]baluma78@gmail.com, [2]pranjana@hindustanuniv.ac.in

***Abstract*—** *An essential part of pharmacovigilance is ADR (Adverse Drug Reaction) detection, which seeks possible drug side effects. Combining Artificial Neural Networks (ANN) with Latent Semantic Analysis (LSA) may be useful for this purpose. ADR detection identifies adverse drug reactions by examining textual data, such as patient reports, social media posts, or medical records. The goal is to categorize text into "ADR" and "non-ADR." The process of extracting and displaying the contextual meaning of words in a corpus of text is referred to as latent semantic analysis or LSA. The dimensionality of the term-document matrix is decreased by applying Singular Value Decomposition (SVD). For ADR detection, ANN models offer several benefits over conventional classifier algorithms such as Logistic Regression (LR), SVM, and Random Forest (RF); they can integrate multi-modal data, manage intricate non-linear relationships, and comprehend textual context. Therefore, we propose an LSA model with an ANN classifier for detecting ADR from text. Results indicate that the LSA with ANN classifier outperforms traditional classifier algorithms like SVM, LR, and RF etc.*

***Index Terms*–** *ADR, ANN, LSA, LR, Random Forest, SVM.*

I. INTRODUCTION

On 29 January, 1848, 15-year-old Hannah Greener passed away while undergoing surgery for an ingrown nail while under chloroform anesthesia, marking the first known instance of an adverse drug reaction. According to the WHO, an adverse drug reaction [16] is a reaction to a toxic and unexpected medication that ensues at dosages normally used in human beings for illness diagnosis [16], or treatment, or prevention [1], or for altering physiological function. It is a serious threat to mankind and the pharma industry. So detecting ADR is an important issue that needs immediate attention. Generally, a lot of clinical trials are conducted to decide drug safety before marketing. It is limited due to duration and cost factors to identify all the possible ADRs. So, after marketing, it is also necessary to detect ADR [2]. Though US Food & Drug Administration Adverse Event Reporting System and WHO Vigibase are authorized sources for the detection of ADR, and the quick growth of social media [14] also contributes to the detection of ADR

Most previous studies suggested that sentiment analysis should be integrated with the detection of ADR. Additionally, data from social media are relatively small compared to FAERS as well as Vigibase. Most ADR cases were unreported during the early manual detection phase. The primary objective of earlier studies was to detect ADR from a sentence. The data referring to ADR mentions from social networking sites were considered. Subsequently, they applied machine learning algorithms to automatically extract ADR mentions. The main limitation encountered in those studies was accuracy.

A more conventional method, LSA captures latent links between concepts by converting text into a lower-dimensional semantic space and focusing on dimensionality reduction. Although LSA can enhance text categorization by detecting latent semantic structures and improving the representation of concepts relevant to ADR, its performance is often inferior to that of ANNs, with an accuracy range of 70-85% on average. Context-dependent interactions, such as nuanced ADR descriptions, are challenging for LSA to capture, making them easier for ANNs to pick up on. Nevertheless, LSA is less computationally intensive and can provide valuable outcomes when working with smaller datasets or in contexts with limited resources. Symptoms, demography, and drug interactions are examples of rarely linear ADR patterns. ANNs model these intricacies using stacked non-linear layers. In contrast to SVM or Logistic Regression, ANNs minimize manual engineering by directly learning pertinent features from raw data or text that has been LSA-transformed. ANNs are effective at processing large datasets, such as millions of social media postings or EHRs. ANNs [15] utilize regularization to filter out irrelevant patterns after LSA reduces noise in text input.

Artificial Neural Networks generally called ANNs [15] are used to normally detect adverse drug reactions (ADRs). To comprehend the "black-box" nature of ANNs, techniques like SHapley Additive exPlanations

(SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are essential. By assisting clinicians, researchers, and regulators in learning the model's prediction of any adverse drug reaction, these tools promote transparency, trust, and adherence to medical standards.

ANNs, especially deep learning models like RNNs, CNNs, and transformers, are highly effective in handling large and complex datasets, automatically learning intricate patterns from unstructured text data directly without going for feature extraction manually. This makes ANNs particularly strong in capturing contextual information, which is crucial for understanding nuanced ADR-related discussions in clinical records, social media posts, or research papers. They typically deliver higher accuracy (85-95%) in ADR detection, especially when dealing with large volumes of data, but require significant computational resources and large annotated datasets for training. The table below shows the comparison between the traditional classifier and ANN

Table 1 Comparison between Traditional Classifier VS ANN

Features	Traditional Classifiers	ANN
Accuracy	Moderate	High
Feature Engineering	Manual	Automatic
Computational Cost	Low	High
Context Understanding	Weak	Strong
Training Time	Fast	Slow
Handling Noise	Poor	Better
Scability	Moderate	High

II. LITERATURE REVIEW

Ahmed Adil Nafea, Nazlia Omar, and Mohammed M. AL-Ani [4] in their work comprehensively analyze the different methods for identifying adverse drug responses (ADRs). To enhance ADR detection, the study addresses biomedical text mining, information extraction methods, and machine learning (ML) [4] as well as deep learning (DL) [3] approaches. It draws attention to the difficulties in ADR detection and suggests ways forward, such as improving detection models through transfer learning and active learning. It also emphasizes how important it is to gather data regarding adverse drug reactions.

T.Zhang et. Al focuses on adversarial transfer learning, a machine learning approach in which a model is always trained to perform well not just on the source domain but also on a target domain by making the features [5] domain-invariant. This would be useful for ADR detection from social media data, which can be noisy and different from traditional clinical data.

D. Tsujii, et al explore the use of Latent Semantic Analysis (LSA) in the extraction of ADR-related information from clinical texts. While LSA was a powerful tool for extracting semantic meaning [6] from text, the study discusses its limitations in handling complex clinical narratives compared to more advanced machine learning models like neural networks

A. Sarker and G. Gonzalez presented Hybrid Semantic Analysis (HSA), a modular natural language processing pipeline, in their research as a means of standardizing [7] social media references of ADRs to conventional medical terminology. The method maps user-generated mentions to Unified Medical Language [18] System concepts by combining rule-based and semantic matching methods, such as LSA. The approach outperformed baselines based only on LSA or MetaMap, attaining an F-measure [18] of 0.62.

Cao Xiao et.al present a three-layer hierarchical Bayesian model that makes use of Latent Dirichlet Allocation (LDA) [8] to forecast ADRs,. The model learns themes that might indicate biological mechanisms connecting ADRs with drug structures by reading each ADR as a symbolic word. The method outperformed other alternatives due to prediction efficiency and accuracy.

Xiaoyi Chen et al assess how well text mining technologies function when used on social media text data in order to identify ADRs [9] and presented the method designed to extract ADR from French social media in this research. A thorough sample size calculation and evaluation corpus constitution are part of the evaluation methodology that is highlighted with an F-measure [18] score of 0.7 for ADR detection

and 0.940 and 0.810 [18] for drug and symptom

III. METHODOLOGY

Dataset Used:

This corpus (CADECv2) consists of medical forum discussions with annotations about adverse drug events (ADEs) reported by patients. It can benefit research on information extraction, or more generally, text mining from social media to identify potential adverse drug responses based on firsthand patient reports. It includes medication or drug categories, and the text represents a user review of their experience with medication, which is presented in sentence form and informal language [13]. First, we will preprocess the data by applying tokenization, stop word removal, and stemming, using spaCy. We have 3548 text files after preprocessing. The output contains key information after reducing redundancy.

After preprocessing, LSA is applied to handle topic modeling and dimension insensitivity [5]. It extracts latent features from the text by reducing the number of terms to identify semantic patterns. However, LSA has the following limitations.

- Choosing the right number of dimensions for SVD in LSA is an issue
- LSA depends on word co-occurrence patterns but does not fully capture the order of words or nuanced meanings.
- LSA might struggle with synonym and polysemy issues
- Performing SVD on large medical corpora is expensive and it lacks real-world knowledge.

IV. MODEL BUILDING WITH ANN

For applications like ADR detection, building an Artificial Neural Network (ANN) requires creating a multi-layered architecture that can recognize intricate patterns in data. To determine the dimensionality of the feature space (such as the reduced LSA components from text data), the first step in the procedure is to define the input layer. Later hidden layers allow the model to capture hierarchical feature interactions and non-linear correlations. These layers are made up of neurons with activation functions such as Tanh or ReLU. For example, in ADR detection, these layers may use the LSA-transformed text representations to learn semantic correlations between adverse effects and drug-related phrases. The output layer produces probabilistic predictions, usually with the use of a softmax (multi-class) or sigmoid (binary classification) activation. Gradient-based techniques like Adam are used to optimize a loss function (e.g., binary cross-entropy [17] for ADR/non-ADR classification) during ANN training, while regularization strategies like dropout or L2 regularization reduce overfitting, which is particularly important when working with noisy or unbalanced pharmacovigilance data. To maintain a balance between model complexity and generalization, hyperparameter tuning—changing the number of hidden layers, the neurons, and the learning rate, along with batch size—is crucial. Robustness is ensured by validation metrics including precision, recall, and AUC-ROC, especially considering the significant consequences of missing ADRs in the medical field. In contrast to conventional models, artificial neural networks (ANNs) [11] are superior at automatically improving feature representations during training, decreasing the need for human feature engineering, and scaling well with big, high-dimensional datasets. Their computational requirements and "black-box" character, however, emphasize the necessity of meticulous design and validation in practical applications.

V. RESULTS AND DISCUSSION

LSA is applied to the data set using the following steps

1. Convert processed data into TF-IDF
2. Apply SVD for LSA
3. Extract topics and top keywords
4. Finally classification

A. Convert Processed data to TF-IDF

Use TF_IDF vectorizer to transform processed data (text) into a matrix limited to a vocabulary of 5000 most relevant words for improving efficiency. The matrix row represents documents and the column represents terms with values showing how important each word is in each document

B. Apply SVD for LSA

LSA identifies hidden semantic structures (topics) by transforming the high-dimensional TF-IDF matrix into a lower-dimensional representation. Using ten components (topics) from truncated singular value decomposition (SVD).

The TF-IDF matrix is broken down into three matrices:
a) The Document-Topic Matrix, [10] or U, illustrates the connections between each document and its subjects.
b) Σ (Singular Values): Indicates the significance of each topic.
c) Important terms for each topic are shown in the V (Topic-Term Matrix).

C. Extract Topics and Keywords

After decomposition, LSA identified the following key topics

1. Topic 1: day, pain, feel, effect, side, drug, work
2. Topic 2: pain, nexium, chest, stomach, severe, reflux, acid, cramp, day, leg
3. Topic 3: day, appetite, sleep, nausea, weight, feel, dry, headache, sweat
4. Topic 4: pain, weight, gain, chest, depression, help, anxiety, leg, attack, muscle
5. Topic 5: nexium, day, cymbalta, weight, month, pain, week, loss, work, gain
6. Topic 6: weight, effect, side, nexium, gain, year, appetite, pound, dry
7. Topic 7: pain, work, sweat, effect, help, side, constipation, day, severe, mouth
8. Topic 8: drug, diarrhea, stomach, weight, gain, never, severe, recommend, eat, heartburn
9. Topic 9: help, nexium, depression, anxiety, appetite, nausea, great, symptom, heartburn
10. Topic 10: sweat, sleep, night, nexium, excessive, drug, bad, mg, muscle, constipation

Table 2 Label Distribution VS No.of Documents

Label Distribution	number of documents
side effect	1818
pain	786
anxiety	166
depression	134

The table above shows the comparison between label distribution and the number of documents with such labels.

D. Classification

The dataset used in this model does not have a separate label file. So following steps are used by the classifiers

- Scan each file for label-related terms
- Assign labels based on detected keywords
- Balance dataset to avoid Bias

Train the classifier and evaluate performance measures like accuracy (A), precision (P), recall(R), and F1-score. The earlier models used machine learning classifiers like SVM, RF, and LR classifier, whereas the proposed model uses ANN classifier due to the following reasons

- Increasing the number of hidden layers for better feature extraction
- Adjusting learning rate to improve convergence
- Training for more iterations to avoid overfitting

Table 3 Performance metrics of various Classifiers

classifier	accuracy	precision	recall	F1-score
SVM	83.2	83.6	83.2	81.8
RF	76.6	79.9	76.6	73.5
LR	77.7	80.2	77.7	74.3
ANN	87.9	87.6	87.9	85.5

The table above highlights the performance measures of various machine learning [12] classifiers like SVM, RF, LR and ANN classifiers. The result shows that ANN always outperforms other classifiers.

CONCLUSION

ADR detection can be significantly improved when LSA is combined with an ANN classifier. This is an important step in enhancing the overall well-being of a patient. While traditional classifiers perform well under constraints of interpretability, processing resources, and labeled data, ANNs excel in handling large-scale ADR detection tasks, capturing contextual meaning and increasing accuracy. The proposed model of LSA with the ANN classifier improves ADR detection compared to other classifiers. The F1-score indicates the effectiveness of the ANN classifier in ADR detection. The drawback can be addressed with the introduction of Deep Learning in ADR detection

REFERENCES

[1] I. R. Edwards and J. K. Aronson, "Adverse drug reactions: definitions, diagnosis, and management," The Lancet, vol. 356, Oct. 2000
[2] T. Zhang et al., "Adversarial neural network with sentiment-aware attention for detecting adverse drug reactions," J. Biomed. Inform., 2021.
[3] R. Damaševičius, R. Valys, and M. Woźniak, "Intelligent tagging of online texts using fuzzy logic," in IEEE Symp. Ser. Comput. Intell. (SSCI), 2016.
[4] A. A. Nafea, N. Omar, and M. M. AL-Ani, "Adverse drug reaction detection using latent semantic analysis," J. Comput. Sci., vol. 17, no. 10, pp. 960–970, 2021.
[5] T. Zhang et al., "Identifying adverse drug reaction entities from social media with adversarial transfer learning model," Neurocomputing, vol. 453, pp. 254–262, Sep. 2021.
[6] J. Tsujii et al., "Latent semantic analysis for adverse drug reaction extraction from clinical text," in Proc. 2006 Workshop on BioNLP, 2006.
[7] A. Sarker and G. Gonzalez, "Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology," J. Biomed. Inform., 2018.
[8] C. Xiao, P. Zhang, W. Chaovalitwongse, J. Hu, and F. Wang, "Adverse drug reaction prediction with symbolic latent Dirichlet allocation," in Proc. AAAI Conf. Artif. Intell., vol. 31, no. 1, 2017.
[9] X. Chen et al., "Mining adverse drug reactions in social media with named entity recognition and semantic methods," Stud. Health Technol. Inform., vol. 245, pp. 322–326, 2017.
[10]A. Adil, N. Nazlia, Z. Omar, and M. Al-qfail, "Artificial neural network and latent semantic analysis for adverse drug reaction (ADR) detection," Baghdad Sci. J., 2023.
[11]A. Rodionov, "Application of neural networks for analysis of sides behavior in the ADR processes," Uzbek J. Law Digit. Policy, vol. 3, no. 1, pp. 21–34, 2025.
[12]S. S. Gulyamov, R. A. Fayziev, A. A. Rodionov, and G. A. Jakupov, "Leveraging semantic analysis in machine learning for addressing unstructured challenges in education," in Proc. 3rd Int. Conf. Technol. Enhanced Learn. Higher Educ. (TELE), Lipetsk, Russia, 2023, pp. 5–7.
[13]A. Cocos, A. G. Fiks, and A. J. Masino, "Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts," J. Am. Med. Inform. Assoc., vol. 24, no. 4, pp. 813–821, 2017.
[14]H. J. Dai and C. K. Wang, "Classifying the adverse drug reactions from imbalanced Twitter data," Int. J. Med. Inform., vol. 129, pp. 122–132, 2019.
[15]S. Shankar et al., "Predicting the adverse drug reaction of two drug contradictions using structural and transcriptomic drug representations to train an artificial neural network," bioRxiv, preprint, 2020.
[16]I. de G. Anderson et al., "Adverse events in a psychiatric hospitalization unit," 2021.
[17]L. Raithel, Cross-lingual Information Extraction for the Assessment and Prevention of Adverse Drug Reactions, Ph.D. dissertation, Technische Universität Berlin, Germany, 2024.
[18]A. Holzinger et al., Machine Learning for Health Informatics: State-of-the-Art and Future Challenges, IOS Press, 2016.