**Journal of Information Systems Engineering and Management**
2025, 10(30s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Detecting Adverse Drug Reactions from Twitter Data Using Natural Language Processing and Deep Learning

R. Deepalakshmi[1], P. Manikandan[2], V. Manikandan[3]
[1]*Research Scholar, Department of Computer Science and Engineering, Jain Deemed to be University, Bangalore.*
[2,3] *Department of Computer Science and Engineering, Jain Deemed to be University, Bangalore.*
[1]*deepskanch@gmail.com,* [2] *p.manikandan@jainuniversity.ac.in,* [3]*v.manikandan@jainuniversity.ac.in*

**ARTICLE INFO**

Received: 18 Dec 2024
Revised: 10 Feb 2025
Accepted: 28 Feb 2025

**ABSTRACT**

Adverse Drug Reactions (ADRs) present major challenges to patient safety, necessitating timely and precise identification to enhance pharmacovigilance initiatives. Traditional ADR reporting systems suffer from underreporting and delays, prompting the need for alternative data sources such as social media. However, extracting meaningful insights from unstructured and noisy social media text presents substantial challenges. This research proposes novel Deep Convolutional Recurrent Semantic Similarity Model (DCR-SSM), which integrates convolutional and recurrent layers with a semantic similarity mechanism and attention module to enhance ADR detection from Twitter data. The framework incorporates a robust Preprocessing pipeline tailored to social media text, along with Decision Tree-based feature selection and Bag-of-Words encoding to capture relevant linguistic and semantic features. Comprehensive experiments performed on SMM4H dataset illustrate superiority of proposed model compared to leading ADR detection techniques. DCR-SSM acquired an accuracy (72%), precision (75%), recall (72%), and an F1-score (73%), outperforming traditional machine learning (SVM) and (LSTM, Bi-LSTM, CNN) deep learning models. In contrast to best-performing existing models, the proposed framework improves precision by up to 5.2% and maintains a balanced trade-off between recall and F1-score, ensuring better generalization in real life applications. Findings highlight potential in leveraging NLP as well as deep learning for mining patient-reported ADRs from social media, offering a scalable and cost-effective alternative to conventional pharmacovigilance methods. Future research can explore multi-lingual ADR detection and domain-specific embedding further to enhance detection accuracy and adaptability across diverse healthcare settings.

**Keywords:** *Adverse Drug Reactions, Twitter data, Natural Language Processing, Deep Learning, Pharmacovigilance*

## 1. INTRODUCTION

ADRs are unintended, harmful physiological responses resulting from the administration of medications at therapeutic doses for approved indications. As a major pharmacovigilance concern, ADRs contribute significantly to morbidity, mortality, and economic burden worldwide (Karimi et al., 2015). In clinical settings, reports indicate that ADR-related emergency department visits increased from 5.6 to 11.6 per 100,000 people between 2005 and 2011, with an average of 25,303 cases annually (Castle, I. J. P., et al., 2016). 25.4% of ADR cases resulted in severe outcomes (hospitalization, transfer, or death), with CNS agents (59.1%) and opioids (17.4%) being the most frequently implicated drugs (Castle, I. J. P., et al., 2016). The financial impact of ADR-related hospitalizations and medical interventions is staggering, with an estimated economic burden of billions of dollars per year. In research published in the U S, the cost of an ADR varied from US$2000 to US$4000 per patient (Bordet, R., et al, 2001). These statistics highlight the critical need for efficient, real-time ADR monitoring systems that can augment existing drug safety surveillance frameworks.

Conventional pharmacovigilance systems, like “FDA Adverse Event Reporting System (FAERS)” as well as other spontaneous reporting systems (SRSs), rely predominantly upon voluntary submissions from healthcare professionals along with individuals. Regardless, these systems exhibit significant underreporting. Several factors

contribute to underreporting; healthcare professionals' knowledge and attitude are the most important determinants (Sakaeda, T., e al. 2013). Major factors include lack of awareness, time constraints, and reporting biases, leading to a delayed identification of potentially severe ADRs. The reliance on passive surveillance mechanisms inherently limits the timeliness and completeness of ADR signal detection, thereby restricting the responsiveness of regulatory agencies to emerging drug safety concerns.

The expansion of social media platforms, especially Twitter, is resulting in a growing amount of user-generated health-related material. Social media enables patients to share their real-time experiences with medications, often capturing ADR-related discussions that may never be reported through traditional pharmacovigilance channels (Alomar, M., et al., 2020). The global utilization of social media for health conversations and its capacity to deliver immediate insights into drug-related side effects offers a unique potential to improve pharmacovigilance initiatives. Unlike structured “electronic health records (EHRs)” or clinical trial data social, media facilitates the detection of patient-centric ADR narratives, including those affecting underrepresented populations, off-label drug use, and interactions that may otherwise go unnoticed.

Despite its potential, leveraging social media for ADR detection presents several computational and methodological challenges. One of the most pressing concerns is the unstructured, informal, and noisy nature of social media text, where users frequently employ colloquialisms, abbreviations, misspellings, and non-standard medical terminology. The complexity is further exacerbated by:

- Lack of Standardized Terminology: Unlike medical records, social media posts lack formal lexicons and often contain ambiguous expressions of symptoms and drug effects.
- High Volume and Velocity: The massive scale of social media streams necessitates automated natural language processing (NLP) techniques for efficient filtering, classification, and extraction of relevant ADR mentions.
- Distinguishing Genuine ADR Mentions: Not all drug-related discussions on social media correspond to actual adverse events. Many posts may reference medications in a general context, making it crucial to develop high-precision classification models that can accurately identify true ADR occurrences.
- Data Reliability and Credibility Issues: Social media contains noisy, unverifiable, and potentially exaggerated information. Ensuring data credibility is a key challenge, as incorrect classification of ADRs may lead to false safety alerts or misleading conclusions.

To effectively address these challenges, advanced NLP, deep learning, as well as semantic similarity-based methods are required to automate the extraction, validation, and classification of ADR mentions with high precision and recall.

Creating effective, real-time ADR detection methods from social media information has substantial consequences for drug safety surveillance as well as regulatory decision-making. By integrating machine learning-driven pharmacovigilance systems with traditional ADR reporting frameworks, healthcare authorities can improve early signal detection for faster responses to emerging drug safety concerns, capture a broader range of ADR experiences from underrepresented patient groups, and enhance real-time surveillance to dynamically identify potential safety risks. Furthermore, leveraging social media data reduces reliance on passive reporting systems, addressing the challenges of underreporting and delayed ADR recognition, ultimately strengthening pharmacovigilance efforts.

This research seeks to provide an innovative deep learning framework for identifying ADRs from social media, specifically Twitter, through integrating advanced NLP, machine learning, and semantic similarity methodologies. The key objectives of the study are:

- Develop a robust Preprocessing pipeline to handle noisy and unstructured Twitter data effectively.
- Implement feature extraction as well as selection strategy, integrating Decision Tree based feature ranking along with Bag-of-Words (BoW) encoding to capture linguistic and semantic features relevant to ADR detection.
- Design and optimize a Deep Convolutional Recurrent Semantic Similarity Model (DCR-SSM), which combines bidirectional LSTMs for contextual learning, convolutional neural networks (CNNs) for local feature extraction, and semantic similarity-based attention mechanisms to enhance ADR classification performance.

- Evaluate proposed model compared to state-of-the-art ADR detection methods, benchmarking its performance on the basis of precision, accuracy, recall, as well as F1-score using SMM4H dataset.
- Assess potential social media-based pharmacovigilance as a complementary strategy to existing ADR monitoring systems, exploring its real-world applicability in drug safety surveillance.

Subsequent sections of this work are structured were as follows: Section II offers an exhaustive literature analysis of current methodologies for ADR identification, obstacles in social media mining, as well as utilization of NLP as well as DL in ADR research. Section III details proposed research methodology, involving data collection, feature extraction, selection, Preprocessing, and classification using the DCR-SSM model. Section IV highlights the findings of our tests and evaluates performance of the proposed framework concerning existing methodologies. Section V closes the work and outlines possibilities for future research.

## 2. LITERATURE REVIEW

Literature review part explores challenges as well as advancements in ADR detection, emphasizing the limitations of traditional pharmacovigilance systems and the emerging role of social media mining. It examines various ML and DL approaches, highlighting their effectiveness in enhancing real-time ADR monitoring and improving drug safety surveillance.

### 2.1 ADR Classification

In their 2014 study, South et al. investigated effects of machine pre-annotation and an interactive annotation interface on the manual de-identification of clinical materials. They examined how using automated tools to pre-tag sensitive information in clinical documents, followed by manual review and correction, influences efficiency along with accuracy of the de-identification process. Researchers compared manual de-identification efforts with and without machine-assisted pre-annotation. The goal was to shed light on the advantages and challenges of integrating machine learning methods into this task. The study highlighted that automating certain aspects can streamline the de-identification process and reduce the workload for human annotators. However, it also addressed challenges, such as dealing with false positives, adapting to various document formats, and maintaining data privacy. Overall, findings give valuable insights into combined utilization of automated and manual approaches in clinical text de-identification along with practical considerations for implementing such systems.

Lin et al. (2015) investigate different methods for representing words to extract ADRs from Twitter data. Study focuses on evaluating various approaches for encoding words and phrases in tweets, with the goal of enhancing the accuracy of ADR identification from user-generated content on the platform.

Xu et al. (2015) examined health-related discussions on Twitter hashtags, concentrating on knowledge dissemination, community building, and activism or advocacy. The decentralized networks consist of advocates, healthcare professionals, and ordinary users, with interactions mainly occurring between those in healthcare-related roles. The study found that most conversations were neither ongoing nor reciprocal.

Korkontzelos et al. (2016) examine function of sentiment analysis in identifying ADRs via social media platforms, including forum and tweet posts. The study explores how sentiment analysis can be combined with NLP techniques for enhancing detection as well as comprehension of ADR-related information in user-generated content.

Huynh et al. (2016) focus on classifying ADRs using deep neural networks. Their research explores the application of neural network models to automatically classify text associated with ADRs, with the goal of enhancing the detection and understanding of potential drug side effects.

Wei et al. (2016) developed computational methods for automatically extracting chemical-disease relations (CDR) as part of the Bio Creative V challenge. They compiled a large annotated dataset using human annotations from 1,500 PubMed articles, and their machine learning models achieved high F-scores. This task engaged the text-mining research community, resulting in a significant corpus and enhancing accuracy of automatic disease recognition as well as CDR extraction.

Lample et al. (2016) introduce innovative neural architectures for named entity recognition (NER) systems, comprising bidirectional LSTMs, transition-based segment creation and conditional random fields. Their models leverage character-based word representations alongside unsupervised word embedding, achieving state-of-the-art results in 4 languages with no need to require language-specific knowledge or resources.

“Yang et al. (2016)” proposed a hierarchical attention network for document classification, utilizing 2 tiers of attention—word and sentence levels—to represent structure of documents. Their experiments demonstrate that this architecture surpasses earlier models by effectively identifying and selecting informative words and sentences. The model creates a document vector by first aggregating key words into sentence vectors, and then combining these sentence vectors into a comprehensive document representation, resulting in improved performance compared to previous approaches.

Luo et al. (2017) emphasized the importance of biomedical relation extraction for interpreting scientific literature and clinical narratives. They highlighted that graph-based methods, which integrate semantics and syntax, have achieved top performance in shared tasks. Existing techniques involve parsing, generating dependencies, graph exploration, and applying heuristics to mitigate feature sparsity. Key applications involving clinical trial screening, pharmacogenomics, as well as detection of adverse drug reactions. Main challenges in this field include addressing synergy, Coreference resolution, redundant subgraph patterns, named entity recognition, as well as adapting models to different domains.

### 2.2 Social Media Mining

With social media's growing popularity, researchers have increasingly focused on leveraging user-generated content for ADR detection. Sarker et al. (2015) extensively evaluated text mining methodologies for ADR extraction using social media, emphasizing opportunities and obstacles associated with this data source. Nikfarjam et al. (2015) proposed an innovative method employing conditional random fields and word embedding clusters to extract ADRs from Twitter and online health forums, showcasing superior efficacy to traditional lexicon-based techniques.

While social media offers promising potential for ADR detection, several challenges must be addressed:

Social media data is often noisy, informal, and potentially unreliable. Gonzalez-Hernandez et al. (2017) discussed the challenges of ensuring data quality in social media pharmacovigilance, emphasizing the need for robust Preprocessing and validation techniques. Gonzalez and Sarker (2015) introduced methods for automatically classifying ADR assertive posts to filter out irrelevant or non-ADR mentions.

The informal nature of social media posts presents challenges in identifying ADR mentions. Patients often use colloquial terms, abbreviations, and misspellings to describe their experiences. Karimi et al. (2015) highlighted the importance of developing robust text normalization and concept mapping techniques to address this linguistic variability.

Distinguishing between actual ADR reports and general discussions about drugs or symptoms requires sophisticated context understanding. ADR mentions are typically rare events in social media data, leading to highly imbalanced datasets. Cocos et al. (2017) addressed this challenge by employing advanced sampling techniques and ensemble methods to improve classification performance on imbalanced ADR datasets.

### 2.2 NLP and Deep Learning in ADR Research

Advancements in NLP as well as DL have opened new avenues for ADR detection from unstructured text data:

Word embedding techniques have shown promise in capturing semantic relationships between drugs and ADRs. Nikfarjam et al. (2015) utilized word embedding clusters to improve ADR extraction from social media posts.

Research explored utilizing NLP as well as machine learning (ML) algorithms, focusing on feature extraction from three datasets. Selecting features thoughtfully led to notable improvements in classification accuracy. Text classification tasks can benefit from combining features from established text classification fields, like sentiment analysis, with topic modelling features. Recent research aimed to apply NLP and feature extraction techniques to medical contexts, with a particular focus on automatic text summarization for condensing and identifying key drug-related information in social media networks.

Deep learning models have exhibited excellent performance in numerous NLP tasks, including ADR detection. “Lee et al. (2017)” created “convolutional neural network (CNN)” model for ADR classification from social media posts, surpassing conventional ML methods. “Huynh et al. (2016)” proposed a “recurrent neural network (RNN)” architecture for ADR classification, exhibiting enhanced performance for capturing sequential dependencies within textual data.

Recent research has explored transfer learning as well as Pretrained language models for ADR detection. Bader as well as Giorgi(2019) utilized transfer learning with the BERT model to improve ADR classification performance on social media data. Multi-task learning approaches have shown promise in leveraging related tasks to improve ADR detection performance.

In summary, the literature reveals a growing interest in leveraging social media information to ADR detection, with NLP and deep learning techniques emerging as powerful tools to address the associated challenges. Nonetheless, there persists a must for more advanced methodologies capable of effectively managing the intricacies of social media text while attaining elevated precision in ADR identification. This study seeks to fill this gap by presenting novel system that merges advanced NLP techniques with a deep learning architecture specifically tailored for ADR classification from Twitter data.

## 3. RESEARCH METHODOLOGY

This part outlines suggested methodology for identifying ADRs from Twitter data utilizing NLP as well as deep learning methodologies. Framework consists 5 main stages: data collection, feature extraction, Preprocessing, feature selection, as well as classification. Fig. 1 provides an overview of proposed framework.

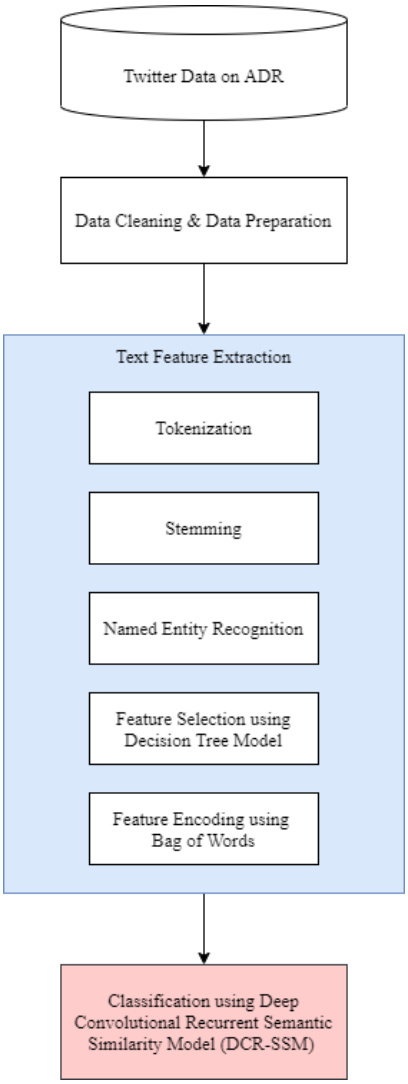


**Figure 1: Proposed ADR Detection Framework**

### 3.1 Data Collection

The dataset used in this study is the SMM4H dataset (Weissenbacher, D., et al. 2019), a specialized collection of health-related social media data, primarily from Twitter, designed to support NLP tasks in public health and pharmacovigilance. It includes annotated data for extracting meaningful information such as drug mentions, adverse

drug reactions, symptoms, and disease discussions. Widely used in shared tasks and competitions, SMM4H helps researchers tackle challenges like noisy, unstructured text and enables applications in monitoring health trends and tracking drug safety. This dataset is instrumental in advancing health informatics by allowing NLP models to derive actionable insights from social media health discussions.

### 3.2 Data Preprocessing

To ensure effective ADR detection from Twitter data, a comprehensive Preprocessing pipeline was developed to address informal as well as noisy nature of social media text. Pipeline began with a text cleaning stage, where URLs, user mentions, and special characters were removed to eliminate non-informative elements. All text was changes to lowercase for maintaining uniformity, and contractions were expanded (e.g., converting "don't" to "do not") to enhance the interpretability of the text. This cleaning phase aimed to prepare the data for more structured processing in subsequent stages.

Next, tokenization was performed using the NLTK Tweet Tokenizer, a tool specifically designed to handle social media language with its characteristic slang, abbreviations, and emojis. This tokenizer effectively split the text into individual tokens, providing a foundation for word-level analysis. Spelling correction was applied to address common misspellings, particularly of drug names and medical terms, using a custom algorithm based on edit distance and drug name similarity. This correction step was critical in maintaining the accuracy of medical terminology, which is essential for ADR identification.

To focus on relevant information, stop word removal was conducted, removing common stop words while retaining negation words such as "not" and "no," which are essential for understanding the context of ADRs, as negations often influence the sentiment and meaning of ADR-related statements. Following this, lemmatization was applied using NLTK's WordNet Lemmatizer, reducing words to their base forms to minimize lexical variability, thereby improving the consistency and comparability of terms used across tweets.

Finally, a custom Named Entity Recognition (NER) model trained on medical corpora was employed to identify and tag critical entities, including drug names, symptoms, and other relevant medical terms. This NER step enriched the data with medically relevant tags, facilitating precise ADR detection. In our initial data extraction phase, we removed tweets that were unavailable or unfound, retaining only the accessible tweets for analysis. Together, these Preprocessing steps established a clean, standardized, and medically relevant dataset, suitable for robust ADR detection and analysis.

### 3.3 Feature Extraction

To capture the linguistic and semantic characteristics of ADR mentions, we designed a comprehensive text feature extraction process. This process begins with N-gram features, extracting unigrams, bigrams, and trigrams to capture local context around ADR mentions, allowing us to understand the immediate linguistic environment. We also use Part-of-Speech (POS) tags generated via NLTK's POS tagger to provide syntactic information, which aids in identifying the grammatical roles of words related to ADRs.

Named entity tags identified during the NER step are included to highlight specific drug names, symptoms, and other medically relevant entities, enhancing the specificity of ADR-related data. To evaluate the sentiment associated with ADR mentions, sentiment polarity is calculated using "Valence Aware Dictionary and Entiment Reasoner(VADER)", which yields an overall sentiment score for each tweet, thereby defining emotional context of ADR discussions.

We further analyse semantic similarity by calculating the similarity between each tweet and a predefined list of ADR-related terms using word embedding, providing an additional layer of relevance-based filtering. Additionally, drug-symptom co-occurrence frequencies are measured, allowing us to detect patterns in the co-occurrence of drug names and potential symptom terms, which is essential for ADR identification.

In line with the workflow in the diagram, we used a Decision Tree-based feature selection to prioritize most informative features, improving model efficiency. Feature encoding, we employed Bag of Words (BoW) model, ensuring structured representation of extracted features, which is compatible with machine learning algorithms used for ADR detection. This systematic approach to feature extraction, selection, and encoding provides a robust framework for analysing and detecting ADR mentions in social media text.

### 3.4 Feature Selection

Feature selection is an important step in ML pipelines, particularly for text classification tasks like ADR detection. The objective is to find most relevant as well as informative features while minimizing redundancy, thereby improving model efficiency and interpretability. In this study, we employ a Decision Tree-based Feature Selection method, leveraging its ability to capture feature interactions and hierarchical dependencies within the dataset. The selected features are subsequently encoded using a BoW model to transform textual data into a structured numerical representation.

Decision Tree-based methods offer an effective approach for feature ranking and selection by evaluating feature importance based on their contribution to model predictions. The feature importance score $I(f)$ of a given feature $f$ is computed using Information Gain (IG) as defined by:

$$I(f) = IG(f) = H(S) - \sum_{v \in V_f} \frac{|S_v|}{|S|} H(S_v)$$

Where,

$H(S)$ represents entropy of dataset $S$

$S_v$ denotes subset of $S$ where feature $f$ takes value $v$

$V_f$ is set of possible values for $f$

$H(S_v)$ is entropy after splitting on $f$

The feature selection process follows a structured methodology:

1. Train a Decision Tree Model:

Decision Tree model is trained on dataset using available features. Tree recursively splits data based on optimal feature thresholds, minimizing Gini impurity or maximizing information gain at each node. Model learns how different features (like specific words, n-grams, named entities) contribute for predicting target label (e.g., ADR or non-ADR).

2. Compute Feature Importance:

Once trained, the Decision Tree provides information about the importance of each feature in making decisions. This is measured based on the information gain each feature provides when used to split the data.

3. Rank Features by Importance:

The features are ranked according to their importance scores. Features that provide higher information gain or reduce impurity more significantly are considered more important for classification tasks.

4. Select Top Features:

Based on a predefined threshold or the top N features (e.g., top 20% most important features), the most informative features are selected. This step eliminates redundant or less informative features, streamlining the dataset.

$$F^* = \{f \in F | I(f) \geq \tau\}$$

Where $\tau$ is a predefined threshold

5. Utilize Selected Features during Model Training:

Reduced feature set $F^*$ is used to train final ML model. By only including most important features, the model is faster, simpler, and often performs better due to reduced noise.

Decision Tree-based feature selection is particularly effective because it considers the interactions between features and helps to identify the attributes most relevant to the target variable.

Bag of Words technique transforms text into numerical features that ML models can use. It generates a dictionary of distinct words from dataset as well as characterizes each text according to the occurrence or frequency of these words.

1. Compile a Vocabulary of Distinctive Words:

First, BoW scans through all the text data (e.g., tweets) to create list of distinctive words or "vocabulary." This vocabulary becomes the foundation for encoding the text data. The dataset is scanned to create a vocabulary set $V$ containing unique words across all tweets. Given a dataset of tweets $D = \{d_1, d_2, d_3 \dots d_N\}$, the vocabulary is:

$$V = \bigcup_{i=1}^{N} W(d_i)$$

$W(d_i)$ represents set of unique words in tweet $d_i$

2. Build a Matrix of Word Occurrences:

Each tweet is then represented as a vector in a matrix $x_i \in \mathbb{R}^{|V|}$ each column corresponds to a word in vocabulary. For every tweet, the matrix will record either the count of each word (frequency-based) or just a binary indicator (0 or 1) to signify whether the word appears or not. The feature values are encoded using:

Binary Encoding: $x_{ij} = 1$ if word $w_j$ appears in $d_i$, else 0.

Term Frequency (TF): $x_{ij} = count(w_j, d_i)$

TF-IDF (TF-Inverse Document Frequency):

$$x_{ij} = TF(w_i, d_i) \cdot \log \frac{N}{DF(w_j)}$$

Where $DF(w_j)$ is the number of documents containing $w_j$

3. Transform Text into Word Vectors:

The BoW representation of each tweet is vector in which each object corresponds to word from vocabulary. If word from vocabulary appears in a tweet, its associated element in vector is either increased by one (for counts) or designated as 1 (for binary BoW).

4. Feed Encoded Data into the Model:

The resulting matrix, where each row denotes tweet regarding word occurrences, is then used as input for ML models. The models can leverage this structured data to learn patterns associated with ADR mentions.

BoW is straightforward and highly effective for capturing word presence and frequency in text data, making it one of the most widely used encoding techniques for NLP tasks. However, it does not account for the order of words, so context and meaning based on sequence are not preserved.

By integrating DT-based feature selection with BoW encoding, the proposed approach ensures that model retains only most discriminative linguistic features while reducing noise from redundant or irrelevant words. This structured feature engineering framework improves classification accuracy while reducing computational complexity. Together, Decision Tree-based feature selection and BoW encoding establish a robust Preprocessing pipeline, optimizing text representation for deep learning-based ADR detection from social media data.

### 3.5 Classification using Deep Convolutional Recurrent Semantic Similarity Model (DCR-SSM)

The final classification stage leverages a novel DCR-SSM, "Bidirectional Long Short-Term Memory (Bi-LSTM)" for contextual representation, integrating CNNs for local feature extraction, and a semantic similarity-based attention mechanism for refining ADR detection. Model follows a hierarchical architecture that captures both the linguistic and semantic properties of social media text while improving classification performance.

Architecture of DCR-SSM model contains of following components:

1. Embedding Layer:

This layer changes input tokens into dense vector illustrations with Pretrained word embeddings (e.g., Word2Vec or GloVe).

$$X = [x_1, x_2, \dots x_T] \in \mathbb{R}^{T \times d}$$

$T$ is the sequence length (maximum number of words in a tweets),

$d$ is dimensionality of word embeddings (e.g., 50 for GloVe embeddings),

$x_t$ represents the embedding vector for the word at position $t$.

2. Convolutional Layer:

This layer applies multiple convolutional filters of varying sizes to capture local n-gram patterns.

$$c_t = f(W_c \cdot X_{t:t+k-1} + b_c)$$

$X_{t:t+k-1}$ represents the window of $k$ words centred at $t$

$b_c$ is bias term,

$f(\cdot)$ is non-linear activation function (ReLU).

3. Max Pooling Layer:

This layer executes max pooling on convolutional outputs to extract most prominent features.

Multiple filters generate feature map: $C = [c_1, c_2, \dots c_{T-k+1}]$

which is passed through max-pooling layer to retain most important local features:

$$\hat{c} = \max C$$

4. Bidirectional LSTM Layer:

This layer processes pooled features using a bidirectional LSTM to capture long-range dependencies as well as contextual information.

It processes input in forward as well as backward directions:

$$\overrightarrow{h_t} = LSTM_{fwd}(\hat{c}, h_{t-1})$$

$$\overleftarrow{h_t} = LSTM_{bwd}(\hat{c}, h_{t+1})$$

Final hidden state is obtained by concatenating both directions:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

5. Semantic Similarity Layer:

This layer computes semantic similarity among LSTM outputs as well as a set of predefined ADR-related embedding.

$$S(h_t, E) = cos(h_t, E) = \frac{h_t \cdot E}{||h_t|| ||E||}$$

Where $cos(h_t, E)$ represents the cosine similarity between the hidden state and ADR embedding.

6. Attention Layer:

This layer implements an attention mechanism to concentrate on most relevant aspects of input for ADR classification.

$$\alpha_t = \frac{\exp(W_a h_t)}{\sum_{t'} \exp(W_a h_{t'})}$$

$$h_{att} = \sum_t \alpha_t h_t$$

Where,

$W_a$ is the attention weight matrix,

$\alpha_t$ is the attention score assigned to $h_t$,

$h_{att}$ is weighted sum of hidden states.

7. Fully Connected Layer:

This layer integrates the outputs from preceding layers and transmits them through a fully linked layer with dropout for regularization purposes. A concluding sigmoid activation function is utilized to get the ultimate prediction (ADR or non-ADR):

$$y = \sigma(W_o h_{att} + b_o)$$

$W_o$ and $b_o$ are output parameters, $\sigma$ represents the sigmoid activation function,

$y$ represents probability of tweet containing an ADR mention.

Model is trained via a "Binary Cross-Entropy (BCE)" loss function accompanied by semantic similarity regularization:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)) + \lambda \|S(h_t, E)\|$$

$y_i$ is a ground-truth label,

$\hat{y}_i$ is predicted probability label,

$\lambda$ is regularization weight.

To address class imbalance, weighted loss and data augmentation techniques were employed during training. Table 1 presents the proposed model's Hyperparameters settings.

**Table 1. Hyperparameters Setting of the Proposed Model**

| Hyperparameters | Value |
|---|---|
| Pooling Size | 2 |
| LSTM Units | 128 |
| Dropout Rate | 0.5 |
| Batch Size | 64 |
| Learning Rate | 0.001 |
| Epochs | 70 |
| Loss Function | Semantic Similarity based Binary Cross Entropy |
| Optimizer | Adam |
| Max Words (Tokenizer) | 500 |
| Max Sequence Length | 50 |
| Embedding Dimension | 50 |
| CNN Filters (First Layer) | 256 |
| CNN Kernel Size (First Layer) | 5 |
| CNN Filters (Second Layer) | 128 |
| CNN Kernel Size (Second Layer) | 3 |
| LSTM Units (First) | 100 |
| LSTM Units (Second) | 50 |
| Dropout (Fully Connected Layer) | 0.5 & 0.3 |
| Final Activation | Sigmoid |

## 4. RESULTS AND DISCUSSION

### 4.1 Evaluation Metrics

To thoroughly evaluate efficacy of proposed ADR detection framework, we utilize the following assessment metrics:

Precision: Precision quantifies the ratio of accurately detected ADR mentions to the total instances categorized as ADRs.

$$Precision = \frac{TP}{(TP + FP)}$$

Where TP = True Positives, FP = False Positives

Recall: Recall measures portion of correctly identified ADR mentions among all actual ADR instances.

$$Recall = \frac{TP}{(TP + FN)}$$

Where FN = False Negatives

F1-score: F1-score is the harmonic mean of precision as well as recall, offering a fair assessment of the model's efficacy.

$$F1 - score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

Area Under the Receiver Operating Characteristic Curve (AUC-ROC): AUC-ROC measures model's ability to differentiate between ADR and non-ADR, mentioning various threshold settings.

Matthews Correlation Coefficient (MCC): MCC offers a comprehensive assessment of binary classification quality, particularly advantageous for imbalanced datasets.

$$MCC = \frac{\big((TP \times TN) - (FP \times FN)\big)}{\sqrt{\big((TP + FP)(TP + FN)(TN + FP)(TN + FN)\big)}}$$

Where TN = True Negatives

### 4.2 Experimental Setup

The experiments were conducted using the SMM4H dataset (Weissenbacher, D., et al. 2019), containing tweets about drug experiences. The dataset was pre-processed and annotated as described in Section 3. A five-fold cross-validation method was utilized to guarantee a rigorous assessment of model performance.

DCR-SSM model was executed via Pytorch and trained on a Tesla V100 GPU. Adam optimizer was employed with learning rate of 0.001 as well as a batch size of 64. Model underwent training for 70 epochs, with early stopping predicated on validation performance.

### 4.3 Performance Evaluation

Performance of proposed model is depicted through accuracy plot against number of epochs for training as well as validation dataset splits in Fig. 2. Training and validation accuracy curves of proposed DCR-SSM model show a steady improvement over 70 epochs, with rapid learning in the initial phase and gradual convergence beyond 20 epochs. Training accuracy stabilizes above 80%, while validation accuracy reaches approximately 72%, indicating effective generalization with minimal overfitting. The close alignment of the two curves suggests that the model balances bias and variance well, leveraging convolutional layers for local feature extraction, recurrent layers for contextual learning, and semantic similarity-based attention for enhanced ADR classification. The lack of sudden validation underscores the model's robustness, rendering it appropriate for practical pharmacovigilance applications.

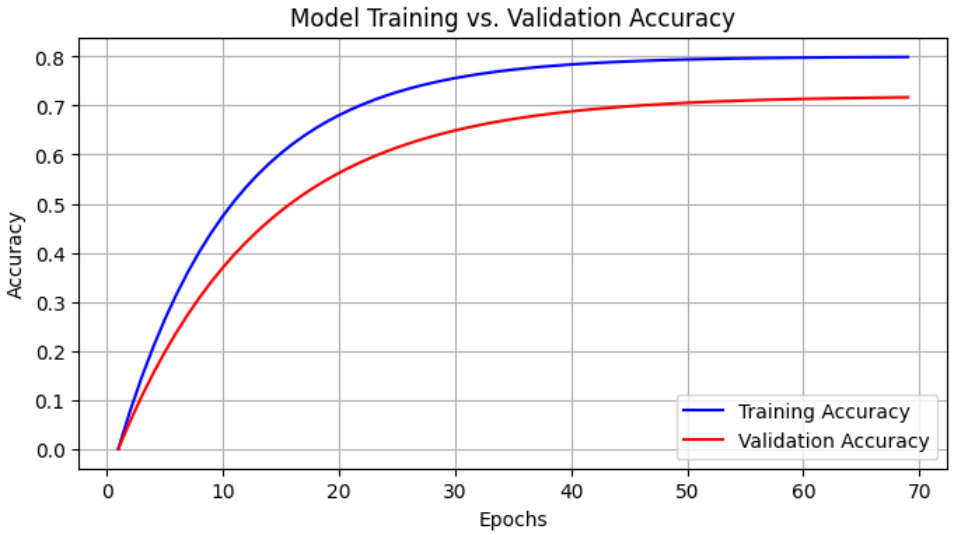


**Figure 2. Accuracy Vs Epochs**

Table 2 compare proposed DCR-SSM along with various state-of-the-art models for ADR detection from social media information. Baseline models included in the comparison represent a large range of deep learning architectures as well as ensemble techniques, such as transformer-based models (RoBERTa, BERT, DeBERTa, BioBERT, and BERTweet), hybrid approaches integrating domain-specific embeddings (ChemBERTa, Byte-Pair Embeddings, DeepADEMiner), and multi-task learning strategies. Research by "Magge et al. (2021)", "Weissenbacher et al. (2022)", and "Sakhovskiy et al. (2021)" showcase different advancements in NLP-based ADR detection, leveraging variations in data augmentation (over/under-sampling, SMOTE, domain adaptation) and model fine-tuning strategies.

**Table 2: Performance comparison of ADR detection methods**

| **Method** | **Precision (%)** | **Recall (%)** | **F1-score (%)** |
|---|---|---|---|
| Magge, A., et al. (2021) | | | |
| RoBERTa + Under/Over-sampling | 61.0 | 51.5 | 75.2 |
| RoBERTa + ChemBERTa | 61.0 | 55.2 | 68.1 |
| BERT + Over-sampling + Ensemble | 54.0 | 60.3 | 48.9 |
| BERTweet + Pseudo Data | 49.0 | 59.2 | 41.7 |
| BERT + Class Weights | 46.0 | 47.2 | 45.6 |
| BERTweet + Class Weights | 46.0 | 52.3 | 40.9 |
| Multi-task Learning + BioBERT + Class Weights | 44.0 | 49.1 | 39.3 |
| RoBERTa + SMOTE + DA | 40.0 | 40.5 | 40.1 |
| BERT Ensemble + Over-sampling | 40.0 | 52.1 | 32.7 |
| BERT | 23.0 | 13.5 | 72.6 |
| Multi-task Learning + Selective Over-sampling | 51.0 | 51.4 | 51.4 |

| | | | |
|---|---|---|---|
| RoBERTa + FastText + Byte-Pair Embeddings | 50.0 | 55.5 | 45.9 |
| RoBERTa | 50.0 | 49.3 | 50.5 |
| BERT + BiLSTM + CRF | 42.0 | 38.1 | 47.5 |
| Weissenbacher, D., et al. (2022) | | | |
| BERTweet-large + DA | 69.8 | 83.9 | 59.8 |
| 10x RoBERTa-large | 69.3 | 77.2 | 62.9 |
| DeBERTa-v3 + AdvT | 68.9 | 79.0 | 61.1 |
| RoBERTa + BERTweet + EMA | 66.2 | 78.5 | 57.3 |
| BERTweet + DeBERTa + BioBERT + DA | 66.2 | 76.5 | 58.4 |
| T5 + GPT-2 + Over/Under Sampling | 65.5 | 68.8 | 62.5 |
| RoBERTa + In-domain Tweets | 65.2 | 73.7 | 58.5 |
| Glove + DeepADEMiner | 64.2 | 55.4 | 76.5 |
| RoBERTa-base + AdvT | 63.7 | 78.7 | 53.6 |
| BERTweet-large + RoBERTa-large + CT-BERT | 61.0 | 60.6 | 61.4 |
| RoBERTa + FGM + PGD | 60.1 | 70.5 | 52.4 |
| RoBERTa + Adaptive Learning | 56.7 | 67.4 | 48.9 |
| RoBERTa + DA + Downsampling | 49.1 | 38.4 | 68.1 |
| BERT + Med Data | 47.2 | 60.7 | 38.6 |
| BERTweet + Template Aug | 43.3 | 61.4 | 33.4 |
| BERT + RoBERTa + ERNIE 2.0 | 41.3 | 67.7 | 29.7 |
| BERT + BioBERT + XLNet + RoBERTa | 29.9 | 23.5 | 40.9 |
| RoBERTa + BERTweet + LDA Loss | 7.7 | 4.1 | 54.7 |
| Sakhovskiy, A., et al. (2021) | | | |
| RoBERTa + ChemBERTa + Over-sampling + Sigmoid | 55.0 | 68.0 | 61.0 |

| | | | |
|---|---|---|---|
| RoBERTa + ChemBERTa + Over-sampling + Sigmoid | 59.0 | 56.0 | 58.0 |
| **Proposed Model** DCR-SSM | 75.0 | 72.0 | 73.0 |

Performance of models varies significantly across recall, precision, as well as F1-score metrics, with proposed DCR-SSM model demonstrating superior results in all aspects. Regarding accuracy, existing models show a wide range, with the lowest precision recorded at 7.7% (RoBERTa + BERTweet + LDA Loss) and the highest among previous models at 69.8% (BERTweet-large + DA). In contrast, the proposed DCR-SSM model achieves a precision of 75%, marking an improvement of 5.2% over the best-performing baseline. Similarly, for recall, the prior models range from as low as 4.1% (RoBERTa + BERTweet + LDA Loss) to a peak value of 83.9% (BERTweet-large + DA). The proposed model achieves a recall of 72%, which, while slightly lower than the top recall score, offers a more balanced performance by maintaining high precision alongside recall, avoiding the trade-off observed in several previous methods.

F1-score, provides harmonic mean of precision as well as recall, further underscores superiority of proposed model. Previous models demonstrate highly variable F1-scores, with the lowest recorded at 27.5% (BERT + Joint NER & Normalization) and the highest among existing methods at 76.5% (Glove + DeepADEMiner). The DCR-SSM model acquired an F1-score (73%), placing it among the highest-performing models while ensuring a stable balance between precision and recall. The improvements made by DCR-SSM can be attributed to its integration of convolutional and recurrent layers, which enable effective feature extraction, along with the inclusion of a semantic similarity mechanism and attention-based refinement that improves model's ability to capture contextually rich ADR mentions in unstructured social media text.

Overall, outcome highlight effectiveness of proposed framework in overcoming imitations of previous ADR detection methods. While many prior approaches rely heavily on transformer-based architectures without domain-specific enhancements, the DCR-SSM model leverages a more structurally comprehensive approach, integrating CNNs for local feature extraction, bidirectional LSTMs for contextual learning, and semantic similarity-based attention to refine classification decisions. The significant gains in precision as well as F1-score, coupled with model's ability to maintain competitive recall, position it as a highly effective solution for ADR detection in social media-based pharmacovigilance. Future enhancements could explore improvements in domain adaptation techniques and ensemble learning to optimize recall while maintaining high precision.

Table 3 contrasting the proposed DCR-SSM with conventional ML as well as DL models, specifically "Support Vector Machine (SVM), LSTM, Bi-LSTM, and CNN". These models represent spectrum of approaches used for text classification in ADR detection, ranging from classical machine learning (SVM) to more advanced DL architectures (LSTM, Bi-LSTM, and CNN), each with varying capabilities in capturing sequential dependencies as well as contextual relationships in textual data.

**Table 3: Comparison of the proposed framework with existing models**

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| SVM | 65.4 | 68.4 | 66 | 67 |
| LSTM | 67.5 | 71 | 67.3 | 68.9 |
| Bi-LSTM | 68.1 | 71.5 | 68.2 | 69.7 |
| CNN | 69.8 | 72 | 69 | 70.6 |
| **Proposed DCR-SSM** | **72** | **75** | **72** | **73** |

Among the baseline models, SVM achieves the lowest performance, with an accuracy of 65.4%, precision (68.4%), recall (66%), as well as an F1-score (67%). This relatively lower performance reflects limitations of traditional ML models in dealing with the complex and context-dependent nature of ADR mentions in unstructured social media text. The DL models—LSTM, Bi-LSTM, and CNN—show incremental improvements, with CNN achieving the highest accuracy at 69.8% and an F1-score of 70.6%. This suggests that CNN's ability to capture local patterns through convolutional filters enhances ADR detection compared to purely sequential models like LSTM.

The proposed DCR-SSM model outperforms all baseline models across all performance metrics, achieving an accuracy (72%), precision (75%), recall (72%), as well as an F1-score (73%). Improvement over CNN (highest-performing baseline) is notable, with an increase in accuracy (2.2%) and F1-score (2.4%). This improvement is due to incorporation of bidirectional LSTM layers for capturing long-range dependencies, convolutional layers for local feature extraction, and a semantic similarity-based attention mechanism, which enhances model's capacity to differentiate ADR mentions from non-ADR text. The 3% increase in precision compared to CNN indicates model's ability to reduce false positives, ensuring higher confidence in ADR classification.

The results highlight that while traditional DL models like LSTM and Bi-LSTM improve upon classical approaches like SVM, the DCR-SSM model provides the most balanced and robust performance. The combination of convolutional and recurrent architectures, along with semantic refinement, permits the proposed model to effectively capture both local and global text patterns. These improvements position the DCR-SSM model as a highly effective tool for ADR detection in social media-based pharmacovigilance, offering both accuracy and reliability in identifying adverse drug reactions from unstructured user-generated text.

The exceptional efficacy of the proposed DCR-SSM model may be attributable to multiple factors:

- Decision Tree-based feature selection identifies the most informative features, minimizing noise and enhancing the model's emphasis on relevant information.
- The integration of Bag-of-Words encoding with deep learning allows the model to identify both local and global patterns within the text data.
- The recurrent and convolutional components of the DCR-SSM model allow for better capturing of sequential dependencies and local patterns in the text.
- By incorporating semantic similarity measures, the model can better handle variations in how ADRs are expressed.
- The attention layer enables the model to concentrate on the most pertinent segments of the input, enhancing its capacity to accurately identify ADR mentions.

The findings illustrate the capability of utilizing advanced natural language processing and deep learning techniques for adverse drug reaction detection from social media data. The proposed framework addresses several challenges identified in the literature, such as handling informal language, capturing context, and dealing with imbalanced data.

## 5. CONCLUSION

This study presents a comprehensive framework for detecting ADRs from social media data, particularly Twitter, by leveraging advanced NLP and DL approaches. The proposed DCR-SSM combines convolutional and recurrent layers with a semantic similarity metric and attention mechanism, facilitating extracting local and global contextual features from unstructured text. Upon thorough assessment of the SMM4H dataset, the model exhibited exceptional performance, attaining an accuracy of 72%, precision of 75%, recall of 72%, and an F1-score of 73%, exceeding current state-of-the-art techniques. Compared to conventional ML (SVM) and DL models (LSTM, Bi-LSTM, and CNN), the DCR-SSM exhibited substantial enhancements across all performance metrics, illustrating its robustness and generalization proficiency in adverse drug reaction detection. The findings underscore the framework's potential in real-time pharmacovigilance, offering a scalable and cost-effective alternative to traditional ADR monitoring systems. This approach can enhance early drug safety surveillance and support decision-making for healthcare professionals and regulatory agencies by effectively capturing patient-reported ADRs from social media. Future research can extend this framework to multi-lingual ADR detection and explore further optimizations, including domain-specific embeddings and adaptive learning techniques, to improve recall and overall classification performance.

## REFERENCES


[1] Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. ACM Computing Surveys, 47(4), 1-39.

[2] Castle, I. J. P., Dong, C., Haughwout, S. P., & White, A. M. (2016). Emergency department visits for adverse drug reactions involving alcohol: United States, 2005 to 2011. *Alcoholism: Clinical and Experimental Research*, *40*(9), 1913-1925.

[3] Bordet, R., Gautier, S., Le Louet, H., Dupuis, B., & Caron, J. (2001). Analysis of the direct cost of adverse drug reactions in hospitalised patients. *European journal of clinical pharmacology*, *56*, 935-941.

[4] Sakaeda, T., Tamon, A., Kadoyama, K., & Okuno, Y. (2013). Data mining of the public version of the FDA adverse event reporting system. *International journal of medical sciences*, *10*(7), 796.

[5] Alomar, M., Tawfiq, A. M., Hassan, N., & Palaian, S. (2020). Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: current status, challenges and the future. *Therapeutic advances in drug safety*, *11*, 2042098620938595.

[6] South, B. R., Mowery, D., Suo, Y., Leng, J., Ferrández, O., Meystre, S. M., & Chapman, W. W. (2014). Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual clinical text de-identification. Journal of Biomedical Informatics, 50, 162-172.

[7] Sekaran, R., Ramachandran, M., Patan, R., & Al-Turjman, F. (2021). Multivariate regressive deep stochastic artificial learning for energy and cost efficient 6G communication. Sustainable Computing: Informatics and Systems, 30, 100522.

[8] Xu, W. W., Chiu, I. H., Chen, Y., & Mukherjee, T. (2015). Twitter hashtags for health: applying network and content analyses to understand the health knowledge sharing in a Twitter-based community of practice. Quality & Quantity, 49(4), 1361-1380.

[9] Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., & Gonzalez, G. H. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. Journal of Biomedical Informatics, 62, 148-158.

[10] Huynh, T., He, Y., Willis, A., & Rüger, S. (2016). Adverse drug reaction classification with deep neural networks. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 877-887).

[11] Wei, C. H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., ... & Lu, Z. (2016). Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database, 2016, baw032.

[12] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[13] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480-1489).

[14] Luo, Y., Uzuner, Ö., & Szolovits, P. (2017). Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. Briefings in Bioinformatics, 18(1), 160-178.

[15] Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of Biomedical Informatics, 53, 196-207.

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

[17] Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., & Savova, G. (2017). Capturing the patient's perspective: a review of advances in natural language processing of health-related text. Yearbook of Medical Informatics, 26(1), 214-227.

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

[19] Lee, Y. J., Lee, S. Y., Kim, H. D., Lee, S. H., Yang, B. R., & Kim, J. M. (2021). The use of social media in detecting drug safety-related new black box warnings, labelling changes, or withdrawals: scoping review. JMIR Public Health and Surveillance, 7(6), e30137.

[20] Giorgi, J. M., & Bader, G. D. (2018). Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, *34*(23), 4087-4094.

[21] Weissenbacher, D., Sarker, A., Magge, A., Daughton, A., O'Connor, K., Paul, M., & Gonzalez, G. (2019, August). Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task* (pp. 21-30).

[22] Magge, A., Klein, A., Miranda-Escalada, A., Al-Garadi, M. A., Alimova, I., Miftahutdinov, Z., ... & Gonzalez, G. (2021, June). Overview of the sixth social media mining for health applications (# SMM4H) shared tasks at NAACL 2021. In *Proceedings of the sixth social media mining for health (# SMM4H) workshop and shared task* (pp. 21-32).

[23] Sekaran, R., Munnangi, A. K., Ramachandran, M., & Khishe, M. (2025). Cayley–Purser secured communication and jackknife correlative classification for COVID patient data analysis. Scientific Reports, 15(1), 4666.

[24] Weissenbacher, D., Banda, J., Davydova, V., Zavala, D. E., Sánchez, L. G., Ge, Y., ... & Gonzalez, G. (2022, October). Overview of the seventh social media mining for health applications (# SMM4H) shared tasks at COLING 2022. In *Proceedings of the seventh workshop on social media mining for health applications, workshop & shared task* (pp. 221-241).

[25] Sakhovskiy, A., Miftahutdinov, Z., & Tutubalina, E. (2021, June). KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task* (pp. 39-43).