# Mini Project — Drug–ADR Signal Detection using NLP & Deep Learning

Student: **Aishwarya Adepu**

https://github.com/AISHwaryaa3012/AI-Driven-Adverse-Event-Signal-Detection-from-Biomedical-Literature-using-NLP.git

## 1. Objective

The objective of this mini project was to develop an automated Drug–ADR (Adverse Drug Reaction) signal detection pipeline. This pipeline integrates clinical statistical analysis (FAERS) with text mining (PubMed) to identify and rank statistically strong drug–event pairs.

The pipeline utilizes the following core components:

- FAERS spontaneous reporting data.
- PubMed biomedical literature.
- NLP-based text mining and Named Entity Recognition (NER).
- Rule-based and similarity-based scoring for evidence.
- Deep-learning dataset preparation (optional step).

The system identifies statistically strong drug–event pairs from FAERS and collects supporting textual evidence from PubMed documents.

## 2. Dataset Description

### FAERS Data (Assignments 1 & 2)

This data provided the statistical basis for signal detection.

- `FAERS_signals.csv`: Contains PRR, ROR, Chi-square, and $p$-values.
- `FAERS_EBGM_results.csv`: Contains Bayesian EBGM, EB05, and EB95 values.
- **Total rows:** 24,069 drug–event pairs.

### PubMed Data

This data provided the textual evidence for signal confirmation.

- `raw_pubmed_fixed_ids.csv`: Contains biomedical abstracts.
- `clean_text.csv`: Cleaned version used for NLP processing.
- **Total documents:** 24,069.

### NER & Extraction Data (Produced Intermediary Files)

- `entities.csv`
- `drug_adr_sentence_candidates.csv`
- `ranked_signals_rulebased.csv`

## 3. Methodology

The overall methodology involved merging statistical data and processing textual data through an NLP pipeline, followed by scoring and ranking.

- **Data Merge:** Merged FAERS statistical outputs (`FAERS_signals` and `FAERS_EBGM_results`).
- **Text Cleaning:** Cleaned PubMed text and split it into sentences.
- **NER:** Applied Named Entity Recognition to extract drugs and ADR terms.
- **Sentence Extraction:** Extracted candidate sentences where both the drug and the ADR appear together.
- **Scoring:** Scored each candidate sentence using rule-based and TF-IDF similarity methods.
- **Aggregation:** Aggregated sentence scores to generate final ranked Drug–ADR signals.
- **Preparation:** Prepared a dataset for deep learning (optional step).
- **Output:** Generated graphs and summary outputs.

## 4. FAERS Statistical Analysis Summary

The FAERS analysis established the baseline statistical risk for each drug-event pair.

| **File** | **Key Metrics Computed** | **Signal Detection Indicator** |
|---|---|---|
| `FAERS_signals.csv` | PRR, ROR values, Chi-Square | Statistical significance, "signal detected" flag |
| `FAERS_EBGM_results.csv` | EBGM, EB05, EB95 | EBGM used to **shrink noisy reports**, EB05 for **conservative detection** |

These two files were merged to create the final statistical dataset: **`combined_faers_all.csv`**.

## 5. PubMed Text Processing

The text data was prepared for accurate NLP processing.

| **Step** | **Details** |
|---|---|
| 1. Cleaning | Lowercasing, removing special characters |
| 2. Structure | Sentence splitting |
| **Output File** | `clean_text.csv` |
| **Total Processed** | 24,069 documents |

## 6. Named Entity Recognition (NER)

The spaCy model was used to identify key clinical terms in the abstracts.

- **NER Model:** spaCy `en_core_web_sm`
- **Entities Extracted:** Chemical/drug names, symptoms, medical events, and ADR-related terminology.
- **Total Extracted Entities:** 4,874
- **All results stored in:** `entities.csv`

## 7. Drug–ADR Sentence Extraction

This stage isolated direct textual evidence linking a drug and its adverse event.

| **Process** | **Details** |
|---|---|
| Candidate Identification | For each FAERS pair (drug, event), check if both appear in the **same sentence**. |
| Data Storage | If found, the sentence is stored with its metadata. |
| Generated File | `drug_adr_sentence_candidates.csv` |
| Total Candidates | 4,951,343 sentences |

## 8. Scoring Approach

Each candidate sentence was scored based on the presence of the entities and the contextual similarity to the signal query.

- ✓ **Rule-based Score:**
  - **3:** Sentence contains both drug & ADR.
  - **1:** Sentence contains only one.
  - **0:** No match.
- ✓ **TF-IDF Similarity Score:**
  - Calculated as the **Cosine similarity** between the sentence vector and the query vector (`<drug> <event>`).
- ✓ Combined Score:

$$CombinedScore = RuleScore + SimilarityScore$$

## 9. Ranking of Drug–ADR Signals

The final Drug-ADR signals were ranked by aggregating the maximum score and counting supporting evidence.

For each drug–event pair, the following were considered:

- Maximum combined score across all supporting sentences.
- Total number of supporting sentences.
- Number of unique PMIDs (PubMed IDs).

**Final Output:** `ranked_signals_rulebased.csv`

### Top Example Signals

| Drug | Event | Score |
|---|---|---|
| DEXTROAMPHET... | Drug ineffective | 4.00 |
| CALCIUM CHLORIDE... | Adverse reaction | 3.995 |
| ... | ... | ... |
| [INSERT TABLE FROM ranked_signals_rulebased.csv] | | |

## 10. Visualizations

The key findings were visualized to highlight the strongest signals.

- **Bar Chart:** Shows the top-ranked Drug–ADR signals by their final combined score.
- [INSERT FIGURE: top_ranked_signals_bar.png]

## 11. Deep Learning Preparation (Optional)

A dataset was prepared for potential future deep learning models.

- **Prepared file:** `dl_sentence_dataset.csv`

| Data Field | Description |
|---|---|
| `sentence text` | The text of the supporting sentence. |
| `drug name` | The extracted drug entity. |
| `event term` | The extracted ADR term. |
| `sentence score` | The combined rule + similarity score. |
| `weak labels` | Labels prepared for training an LSTM/BERT model. |

**Note:** A deep learning model was not trained due to an insufficient positive sample count.

## 12. Final Outputs Generated

All final and intermediary files were organized for submission:

- **Outputs saved in:** `results/` and `figures/`

| **Submitted File** | **Purpose** |
|---|---|
| `ranked_signals_rulebased.csv` | Final ranked list of Drug-ADR signals. |
| `drug_adr_sentence_candidates.csv` | List of all sentences containing the pairs. |
| `entities.csv` | All entities extracted by NER. |
| `combined_faers_all.csv` | Merged FAERS statistical data. |
| `clean_text.csv` | Cleaned PubMed abstracts. |
| `summary.csv` and<br>`report_summary.txt` | Summary files. |
| `figures graphs` | All generated visualizations. |

## 13. Conclusion

A complete automated Drug–ADR signal detection pipeline was successfully implemented. The pipeline effectively integrates FAERS statistical analysis with PubMed text mining, resulting in high-confidence ranked drug–event signals. The combination of NER, sentence extraction, rule-based scoring, and similarity scoring produced strong evidence indicators for predictive pharmacovigilance. The groundwork for deep-learning extension was completed and can be built upon in future work.