**README: Adverse Drug Reaction (ADR) Text Dataset**

**Overview**

This dataset contains text data related to Adverse Drug Reactions (ADR) and Non-Adverse Drug Reactions (Non-ADR). It is divided into two files:

1. **Labeled Data**: Text data that has been manually labeled as either ADR (Adverse Drug Reaction) or Non-ADR.
2. **Unlabeled Data**: Text data that has not been labeled and can be used for various tasks such as clustering, unsupervised learning, or label prediction.

The source of this dataset is the same, and the data in both files have the same structure, although the data in the labeled data file doesn’t appear in the unlabeled data file.

The dataset is intended for research in areas such as natural language processing (NLP), machine learning, and pharmacovigilance, where it is used to identify and classify adverse drug reactions based on text data.

**Dataset Files**

**1. Labeled Data (File: ADR\_Labeled\_Data.csv)**

* **Description**: This file contains text data that has been labeled as either representing an Adverse Drug Reaction (ADR) or a Non-Adverse Drug Reaction (Non-ADR).
* **Format**: CSV file with two columns: text and label.

**Columns:**

* text: A string representing the textual content, which describes a patient’s history, including medical events, drug reactions, or patient feedback.
* label:
  + 1: Represents text classified as ADR (Adverse Drug Reaction).
  + 0: Represents text classified as Non-ADR.

**Sample Data:**

|  |  |
| --- | --- |
| **text** | **label** |
| “she was admitted to the icu for close monitoring while the work-up for presumed hemolytic anemia continued. past medical history: cad (cath done at osh because of ekg changes revealed "mild cad" which was not intervened upon) allergies/adverse reactions” | 1 |
| "No side effects reported after the medication……" | 0 |
| "in summary, the patient is a 51-year-old male who was admitted to the [\*\*hospital1 188\*\*] with what was felt to be an adverse reaction to fentanyl and benadryl administration" | 1 |
| "The drug worked perfectly with no issues, patient was discharged earlier than expected….." | 0 |

**2. Unlabeled Data (File: ADR\_Unlabeled\_Data.csv)**

* **Description**: This file contains text data that has not been labeled. It can be used for unsupervised learning, clustering, or as a dataset for future labeling efforts.
* **Format**: CSV file with one column: text.

**Columns:**

* text: A string representing the textual content, which describes a patient’s history, including medical events, drug reactions, or patient feedback. No labels are provided in this file.

**Sample Data:**

|  |
| --- |
| **text** |
| " he had no adverse reactions to any of his blood product transfusions. he was started on revlimid…." |
| " the patient experienced an adverse reaction during infusion of rituxan ( see below), but tolerated the remainder of her treatment course without complication" |
| "the treatment went well and the patient was responding poasitely to treatment, family members were there to take care of the patient" |
| " pt was on amiodarone in the past, which worked well for her initially but then discontinued its use in [\*\*month (only) 547\*\*] due to side effects " |

**File Details**

**Labeled Data File:**

* **File name**: ADR\_Labeled\_Data.csv
* **Size**: 3.5 MB
* **Number of rows**: 500
* **Columns**:
  + text (string)
  + label (binary: 0 or 1)

**Unlabeled Data File:**

* **File name**: ADR\_Unlabeled\_Data.csv
* **Size**: 20.2 MB
* **Number of rows**: 5526
* **Columns**:
  + text (string)

**Data Source and Collection Method**

The text data in this dataset has been collected from the publicly available MIMIC-III (Medical Information Mart for Intensive Care) database. MIMIC-III is an extensive, single-center database containing information about patients admitted to critical care units at a large tertiary care hospital. We extracted patient reports from this database by considering the most crucial attributes that determine drug reactions and side effects. The labeled data has been manually annotated by healthcare professionals to classify each entry as either an ADR (Adverse Drug Reaction) or a non-ADR.

**Usage**

This dataset can be used for the following tasks:

* Supervised learning tasks such as ADR classification using the labeled data.
* Unsupervised learning or clustering using unlabeled data.
* Training and evaluation of machine learning models for pharmacovigilance.
* NLP tasks like sentiment analysis, text classification, or named entity recognition in the medical domain.