**# CDC VAERS Text Classification Using BERT**

This is a rapid research project to showcase my skills on getting up-to-speed on a project over a single weekend.

Background: The CDC and FDA collect vaccine adverse events through the Vaccine Adverse Event Reporting System (VAERS). There are 2 free text fields where descriptions of the events are captured. Coders search for specific terms in these fields in the form and code them to searchable and consistent MedDRA terms.

Approach: A Fine-Tuned LLM DistilBERT to Classify User VARES Adverse Event Symptoms Text Descriptions To predict more than just the first symptom (SYMPTOM1), a different approach is needed to handle multiple label prediction. This is typically done using a multi-label classification setup, where each symptom is treated as a separate label, and the model learns to predict the presence or absence of each symptom independently (using a MultiLabelBinarizer).

Results:

Next Step:

**Appendix: VAERS – Vaccine Adverse Event Reporting System**

1. Main site - https://vaers.hhs.gov/data/datasets.html
2. Data use - https://vaers.hhs.gov/docs/VAERSDataUseGuide\_November2020.pdf
3. Data collection – passive collection of online or via fillable pdf reports
   1. A close-up of a form

      Description automatically generated
4. MedDRA symptoms
   1. The MedDRA codes provided in the dataset are called the "Preferred Terms"; there are more than 17,000 Preferred Term codes
   2. 2023 1st symptom column has 136143 codes, 4594 unique
   3. Code #18 and #19, the description of the adverse event using MedDRA dictionary
      1. The fields described in this table provide the adverse event coded terms utilizing the MedDRA dictionary. Coders will search for specific terms in Items 18 and 19 in VAERS 2 form or Boxes 7 and 12 on the VAERS 1 form and code them to a searchable and consistent MedDRA term; note that terms are included in the .csv file in alphabetical order.
      2. There can be an unlimited amount of coded terms for a given event. Each row in the .csv will contain up to 5 MedDRA terms per VAERS ID; thus, there could be multiple rows per VAERS ID. For each of the VAERS\_ID’s listed in the VAERSDATA.CSV table, there is a matching record in this file, identified by VAERS\_ID.