**CDC Text Classification Challenge**

**Overview**

Every day, work-related injury records are generated. In order to alleviate the human effort expended with coding such records, the [Centers for Disease Control and Prevention (CDC)](https://www.cdc.gov) [National Institute for Occupational Safety and Health (NIOSH)](https://www.cdc.gov/niosh), in close partnership with the Laboratory for Innovation Science at Harvard (LISH), is interested in improving their NLP/ML model to automatically read injury records and classify them according to the Occupational Injury and Illness Classification System (OIICS).

**Goal**

This [challenge](https://www.topcoder.com/challenges/30103825) represents a classification problem that is expected to be solved using **R**. The solution may be based on some appropriate [DSPA techniques](http://dspa.predictive.space/) and must be provided as self-contained raw Rmd (R notebook source) and HTML (report output).

**Inputs**

* **Training file** is a spreadsheet, with 4 columns (**text**, **sex**, **age**, and **event**). A CSV file with a header includes.

1. **text**. This column describes the raw injury description text data.
2. **sex**. This is a categorical variable, describing the sex of the related person.
3. **age**. This is a positive integer variable, describing the age of the related person.
4. **event**.This is the target variable, specifying the OIICS label to be classified. There are 48 unique labels in total. The [Occupational Injury and Illness Classification System (OIICS)](https://wwwn.cdc.gov/wisards/oiics/) score corresponding to “[Event/Exposure](https://wwwn.cdc.gov/wisards/oiics/Trees/MultiTree.aspx?TreeType=Event)”, Event-Label value range: [1-100].

Event/Exposure: The event or exposure describes the manner in which the injury or illness was produced or inflicted by the source of injury or illness. When the injury or illness occurred as a result of contact with or exposure to an object or substance, select the event or exposure which best describes the manner in which that contact or exposure occurred. The event categories are listed in order of precedence, such that the categories are ordered as follows (see [p.275 of the OIICS manual](https://www.bls.gov/iif/oiics_manual_2010.pdf)):

* 1. Violence and other injuries by persons or animals
  2. Transportation incidents
  3. Fires and explosions
  4. Falls, slips, trips
  5. Exposure to harmful substances or environments
  6. Contact with objects or equipment
  7. Overexertion and bodily reaction

Others: Non-classifiable

When two or more of these events occurred, select the first event listed above. In general, when two or more events occurred within the same subcategory, select the first event listed in the code sequence. Injuries involving the normal worksite tasks and actions of the injured worker and co-workers are considered accidental by default unless information to the contrary is provided.

* **Testing data:** Need to build a model based on the above training data. The model will need to make predictions using a separate **test** file.

The test file is a spreadsheet, with only 3 columns (**text**, **sex**, and **age**). This CSV file contains a header. The format is the same as the training file, but the **event** column will be missing. Once your model is trained, it should be able to consume the test file and produce the prediction file by filling in the Event column. Specifically, your output will be a CSV file with all 4 columns, keeping the same order as the test file.

**Dataset**

There are **229,820** records in total.

* **Training Data**: A random sample of 153,956 records with the outcome **event** column included as **the training set**.
* **Testing Data**: We need to use this training set to develop your model and use the remaining 75,864 of the records are held out for the independent testing the model. The testing data consists of **the same 3-columns that are present in the training set** (The last **event** column is excluded).

The goal is to generate all event predictions for the records in the testing set, using the same order.

**Baseline Model**

Prior reports build a [Bidirectional Encoder Representations from Transformers (BERT) model](https://arxiv.org/abs/1810.04805) using these data achieved an accuracy of around **87%** based on local testing. Can we improve this solution?

**Performance assessment**

Define and utilize assessment functions like **weight F1 score** to measure the performance of this multi-class classification problem. Focusing on the **i**-th OIICS label, we will transform the ground-truth labels and predicted labels to binary labels: **1** means it is the **i**-th OIICS label, while **0** means it is not. In this way, we can build a confusion matrix and then compute its F1 score **F1\_i**.

We have **48 unique OIICS labels in total**. In order to combine those F1 scores of different labels, we use the label frequency in the ground-truth labels as weights. Suppose the **i**-th OIICS label appeared times in the ground-truth labels. We then have the final score as:

The final normalized score: .