**[***Injury Text Classifier***]**



*BSAN 6200: Text Mining and Social Media Analytics*

Author:   
**Eric Wu**

5/6/2020

**Table of Contents**

[Background 1](#_Toc39886986)

[Objective 1](#_Toc39886987)

[Business (or Industry) Overview 1](#_Toc39886988)

[Methodology 1](#_Toc39886989)

[Data 1](#_Toc39886990)

[Process 2](#_Toc39886991)

[Results 3](#_Toc39886992)

[Specifics 3](#_Toc39886993)

[Binary Classifier: 3](#_Toc39886994)

[Multi-class Classifier: 3](#_Toc39886995)

[General Patterns 1](#_Toc39886996)

[Recommendations 5](#_Toc39886997)

[For Practitioners and Stakeholders 5](#_Toc39886998)

# Background

## Objective

Applied Natural Language Processing to the injury narratives from the NEISS data and visualized the results to understand the frequently used words for each kind of injury. On top of that, I want to build a machine learning classifier to classify the injury type based on the text data.

## Business (or Industry) Overview

The data from the National Electronic Injury Surveillance System is raw and unstructured, moreover, it contains a lot of medical abbreviations that a normal person could not easily. Therefore, how could I, as a business analyst, help the CPSC builds an automatic classifier model to classify the injury type instead of typing it manually. Moreover, the product manufacturers, the customers, or even the parents of a kid to know what activities, products, and places they could avoid from being injured.

# Methodology

## Data

[NEISS Data <https://www.cpsc.gov/cgibin/NEISSQuery/home.aspx>]

I collected the data from 2015 to 2018 and I have got 18 variables and 1482899 rows of records. The project mainly focused on the “narrative” and “diagnosis” column. Below is an example of “narrative” column:

'2 YOF, PT WAS CLIMBING ON A BAR STOOL WHEN IT FELL BACKWARDS CAUSING EYEBROW CONTUSION'

After the preprocessing, the “narrative” column will now look like this

'patient climbing bar stool fell backward causing eyebrow contusion'

As you can see, the “narrative” has been transferred to lower cases, and all the unnecessary words and stop words have been removed.

The “diagnosis” column is the ID of the injury type. For example:

|  |
| --- |
| 59 - LACERATION |
| 60 - DENTAL INJURY |

## Process

The main NLP method I used to clean the data is Regular Expression. I first replaced the medical abbreviation to normal word because some of the abbreviations contained punctuations, for example, “d.x.” is the abbreviation of diagnosis. Ideally, I wanted to get rid of all the punctuation to make our model more accurate. However, if I deleted the punctuation before I replaced the words, “d.x.” would be split as “d x ” and it would become more difficult to replace it ever since. Therefore, I decided to remove the punctuation in our second step. Before that, I needed to create a space between each word and the punctuation following it. For example, “he is a boy.” => “he is a boy .”, and now, I could easily remove the punctuations without influencing the words.

The next step is to remove the useless words and stop words. For instance, '2 YOF, PT means “ 2 years old of female patient”. I wanted to get rid of these words because first, there are already columns call “gender” and “age”, and second, these words are misleading when I count the frequency of common words.

Third, after all the above preprocessing, I used the word cloud to visualize our result. The input of our word cloud function is the type of injury. For example, if the user wants to know the most common word for laceration, they could type the corresponding ID of laceration to get the output.

Lastly, I want to build a model to classify the injury type based on the clean text data I have got. I trained a logistic regression model on one balanced data set of category-specific reports (e.g. 10000 reports: 5000 laceration-related vs. 5000 no laceration-related reports) and then test the model on a test data set that contains another set of 10000 injury reports (5000 laceration-mentioned reports vs. 5000 no laceration-mentioned reports). Coming out with this kind of result will help the injury prevention experts at CPSC to know that a machine learning model will rapidly discover a higher percentage of category-specific injury reports than just manually identifying a set of category-specific reports by humans from a very large pool of text data. Moreover, to make our project more comprehensive, I have extracted the top 6 classes of injury and apply a multi-class classifier to test if the text data could classify multiple levels of output as well.

In conclusion, I used Regular Expression to clean the text and remove the punctuation, NLTK toolkit to get rid of the stop words, word cloud to visualize our result, and build a machine learning model to classify the injury type by using the cleaned text data. The total lines of code are above 500 lines.

# Results

## Specifics

### Binary Classifier:

The example target group I am using here is Laceration and the accuracy of the binary logistic model is 0.9776.

The confusion matrix starts with True Negative on the top left, False Negative on the bottom left, False Positive on the top right, and True Positive on the bottom right.

True Negative: Correctly classified the diagnosis that is not laceration.

False Negative: The diagnosis is laceration but has been classified as not laceration.

False Positive: The diagnosis is not laceration but has been classified as a laceration.

True Positive: Correctly classified the diagnosis that is laceration.

0 1

0[[4925 75]

1 [ 149 4851]]

### Multi-class Classifier:

The target classes are the top six injury type mentioned below and the accuracy of the multi-class decision tree model is 0.94458.

The confusion matrix is:

Sequentially, Contusion\_abrasion, fracture, laceration, internal\_injury, strain\_sprain, and others

0 1 2 3 4 5

0[[16119 119 192 155 167 409]

1 [ 183 16897 69 122 122 251]

2 [ 250 97 21082 184 46 243]

3 [ 216 165 210 10217 62 282]

4 [ 204 68 36 37 16201 355]

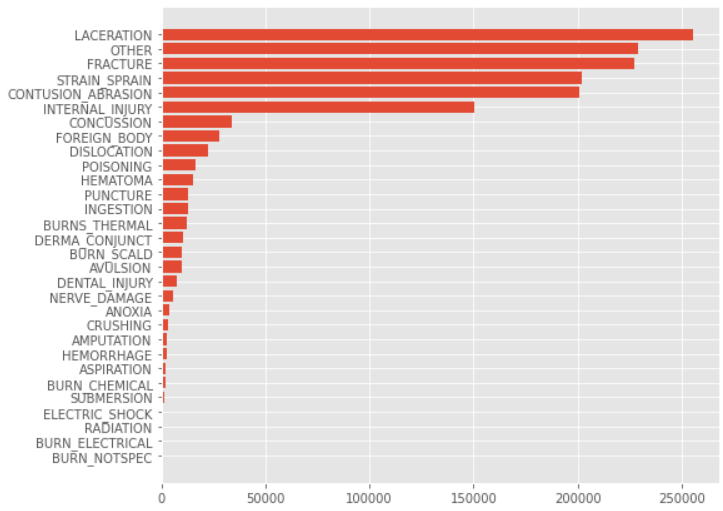
5 [ 430 206 191 191 280 13942]]

There were (16119+119+192+155+167+409) = 17161 points in the contusion\_abrasion (label 0). Out of these, our model was successful in identifying 16119 of those correctly in label 0, but 119 were marked as label 1, 192 were label 2, etc. Similarly look at second row. There were 17644 points in fracture, but 16897 of them were marked correctly. Our classifier predicted 183 in contusion\_abrasion and 69 in laceration, etc.

## General Patterns

As mentioned above, I have 1482899 rows of records. Moreover, after the cleaning step, I have a total of 102892136 words in our corpus.

The frequency of each injury type is shown below:



As I can see, the top six injury type, laceration, other, fracture, strain sprain, contusion abrasion, and internal injury, have accounted for over 85% of the portion.

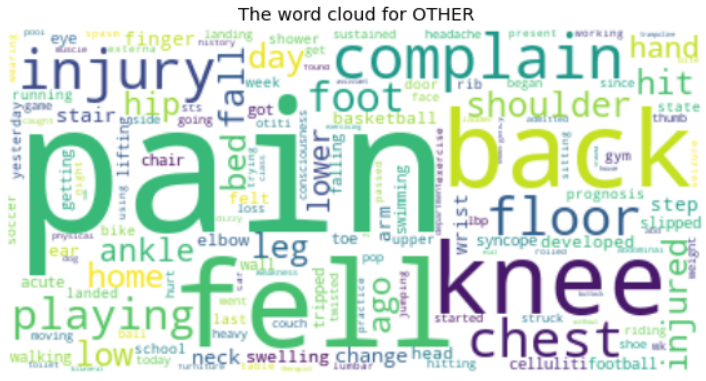
In the next step, I would generate the word cloud for these top injuries to understand the common words in their diagnosis.

Laceration:



The reasons that cause laceration are cutting fingers, felling from the bed, and hitting head and face, etc.

Other:



As you can see, the symptoms for others are mainly knee pain, chest pain, and back pain.

Fracture:



The main reason that caused fracture is because of the sports event.

Strain and Sprain:



Sports events are the main reason that causes strain and sprain as well. Moreover, most of the injured parts are ankle, knee, and lumbar.

Contusion and abrasion:



The patients who are diagnosed contusion is because of falling off the floor, bed, or stair.

Internal Injury:



The internal injury is mainly about head injury, particularly closed head.

# Recommendations

## For Practitioners and Stakeholders

Injury Prevention Expert

* Help the injury prevention experts like CPSC to know that a machine learning model will rapidly discover a higher percentage of category-specific injury reports than just manually identifying a set of category-specific reports by humans from a very large pool of text data

Parent

* Be more cautious when the kids are doing sports events especially basketball and soccer.
* Add the carpet to the floor to avoid serious falling injuries.
* Place the sharpen tools in the place the children could not reach

Hospital

* Put more medical resources to the top 6 injuries
* Let people know the common injury types and teach them how to avoid getting hurt.

Protective clothing

* Focus on the knee and ankle protector to appeal and protect the customers

To conclude, I now have a solid model that could classify and predict the injury type automatically. Moreover, the NLP of cleaning and visualizing millions of words could be applied to different sources of text reviews to do more further analysis.