Preparing the Severe Injury Reports for ML

Abdullah Alsharef

3/20/2020

Contents

1 Introduction

The goal of the this document is to prepare the OSHa Severe Database for ML/NLP analysis. We are going to modify the categories to fit the criteria of the big dimenstionality.

2 Contractor

We'll use level 2 in prediction as an attribute or predictor.

- We'll use level 2.
- NAICS_L4 == "Construction" ~ NA_character_ ,
- NAICS_L4 == "Specialty Trade Contractors" ~ "Other Specialty Trade Contractors",
- NAICS_L4 == "Land Subdivision" ~ "Other Heavy and Civil Engineering Construction"

Discard the two construction cases and do the prediction.

```
load("E:/Research/Safety Research 2020/Severe Injury_OSHADATA/Datasets/R Project/OSHA_Severe/Final_Prep

# filter the needed variables

cons_v3 <- cons_v2 %>%
  select(ID,Final_Narrative, NAICS_L3, NAICS_L4, NAICS_L5,  event_L1,   event_L2,   event_L3,   part_L2

# the contractor variable
cons_cont <- cons_v3 %>%
  select(ID, Final_Narrative, NAICS_L4) %>%
  mutate( Contractor = case_when(
   NAICS_L4  == "Construction" ~ NA_character_ ,
   NAICS_L4  == "Specialty Trade Contractors" ~ "Other Specialty Trade Contractors",
   NAICS_L4  == "Land Subdivision" ~ "Other Heavy and Civil Engineering Construction",
    TRUE ~ as.character(NAICS_L4)))  %>%
  filter(!is.na(Contractor)) %>%
  select(-NAICS_L4)

cons_cont %>%
  count(Contractor) %>%
  mutate(len = length(unique(Contractor)))
```

Contractor	n	len
Building Equipment Contractors	1625	9
Building Finishing Contractors	611	9
Foundation, Structure, and Building Exterior Contractors	2105	9
Highway, Street, and Bridge Construction	703	9
Nonresidential Building Construction	1266	9
Other Heavy and Civil Engineering Construction	257	9
Other Specialty Trade Contractors	752	9
Residential Building Construction	364	9
Utility System Construction	878	9

```
# Output the data set

#write.csv(cons_cont, "Final_Prepetation/Contractor.csv", na = "")
```

3 Event

We'll use Level 1. Exclude all nonclassifiable cases - n = 91 cases.

3.1 V1

```
cons_event <- cons_v3 %>%
  select(ID, Final_Narrative, event_L1) %>%
  mutate(Event = tolower(event_L1) ) %>%
  filter(Event != "nonclassifiable")

cons_event_count <- cons_event %>%
  count(Event, sort = T)


# write.csv(cons_event, "Final_Prepetation/cons_event_1.csv", na = "")
```

##V2

- We'll use level 2 with combining some categories together.
- We'll call general categories "Other . . . ."
- Exclude nonclassifiable

```
cons_event_v2 <- cons_v3 %>%
  select(ID, Final_Narrative, event_L2) %>%
  mutate( Event_v2 = case_when(
   event_L2 == "Fall, slip, trip, unspecified" ~ "Other - falls, slips, trips",
   event_L2 == "Jumps to lower level" ~ "Other - falls, slips, trips",
   event_L2 == "Jumps to lower level" ~ "Other - falls, slips, trips",
   event_L2 == "Slip or trip without fall" ~ "Other - falls, slips, trips",
   event_L2 == "Fall or jump curtailed by personal fall arrest system" ~ "Other - falls, slips, trips",
   event_L2 == "Fall, slip, trip, n.e.c." ~ "Other - falls, slips, trips",
   event_L2 == "Contact with objects and equipment, unspecified" ~ "Other - contact with objects and eq
   event_L2 == "Struck, caught, or crushed in collapsing structure, equipment, or material" ~ "Other -
   event_L2 == "Contact with objects and equipment, n.e.c." ~ "Other - contact with objects and equipme
   event_L2 == "Rubbed or abraded by friction or pressure" ~ "Other - contact with objects and equipmen
   event_L2 == "Pedestrian vehicular incidents" ~ "transportation incidents",
   event_L2 == "Nonroadway incidents involving motorized land vehicles" ~ "transportation incidents",
   event_L2 == "Roadway incidents involving motorized land vehicle" ~ "transportation incidents",
   event_L2 == "Water vehicle incidents" ~ "transportation incidents",
   event_L2 == "Transportation incident, unspecified" ~ "transportation incidents",
   event_L2 == "Animal and other non-motorized vehicle transportation incidents" ~ "transportation inci
   event_L2 == "Aircraft incidents" ~ "transportation incidents",
   event_L2 == "Rail vehicle incidents" ~ "transportation incidents",
   event_L2 == "Fires" ~ "fires and explosions",
```

```
  event_L2 == "Explosions" ~ "fires and explosions",
  event_L2 == "Fire or explosion, unspecified" ~ "fires and explosions",
  event_L2 == "Overexertion involving outside sources" ~ "overexertion and bodily reaction",
  event_L2 == "Other exertions or bodily reactions" ~ "overexertion and bodily reaction",
  event_L2 == "Overexertion and bodily reaction, unspecified" ~ "overexertion and bodily reaction",
  event_L2 == "Bodily conditions, n.e.c." ~ "overexertion and bodily reaction",
  event_L2 == "Animal and insect related incidents" ~ "violence and other injuries by persons or animals
  event_L2 == "Intentional injury by person" ~ "violence and other injuries by persons or animals",
  event_L2 == "Injury by person-unintentional or intent unknown" ~ "violence and other injuries by perso
  event_L2 == "Exposure to other harmful substances" ~ "exposure to harmful substances or environments"
  event_L2 == "Exposure to harmful substances or environments, unspecified" ~ "exposure to harmful subst
  TRUE ~ as.character(event_L2)))  %>%
  filter(event_L2 != "Nonclassifiable")


cons_event_v2_count <- cons_event_v2 %>%
  count(Event_v2, sort = T)


# write.csv(cons_event_v2, "Final_Prepetation/cons_event_v2.csv", na = "")
```

4 Hospitilization

```
hosp <- cons_v3 %>%
  select(ID,Final_Narrative, Hospitalized_Y_N)

#write.csv(hosp, "Final_Prepetation/hosp.csv", na = "")
```

5 Amputation

```
amp <- cons_v3 %>%
  select(ID,Final_Narrative, Amputation_Y_N)

# write.csv(hosp, "Final_Prepetation/amp.csv", na = "")
```

6 Source

We'll use level 1. Exlude "Nonclassifiable".

```
cons_Source <- cons_v3 %>%
  select(ID, Final_Narrative, source1_L1) %>%
  mutate(Source = tolower(source1_L1) ) %>%
  filter(Source != "nonclassifiable")

cons_Source %>%
  count(Source, sort = T)
```

Source	n
structures and surfaces	2199
parts and materials	1516
tools, instruments, and equipment	1479
machinery	1224
vehicles	724
other sources	462
containers, furniture and fixtures	204
chemicals and chemical products	176
persons, plants, animals, and minerals	159

```
# write.csv(cons_Source, "Final_Prepetation/cons_Source.csv", na = "")
```

7 Body Part

We'll use level 1. We'll merge "Head" with "Neck".

```
cons_part <- cons_v3 %>%
  select(ID, Final_Narrative, part_L1) %>%
  mutate(Part = tolower(part_L1) ) %>%
  filter(Part != "nonclassifiable") %>%
  mutate(Part_v2 = case_when(
    Part == "head" ~ "head and neck",
    Part == "neck, including throat" ~ "head and neck",
    TRUE ~ as.character(Part))) %>%
  select(-Part,-part_L1)

# write.csv(cons_part, "Final_Prepetation/cons_part.csv", na = "")
```

8 Nature

we'll use level 2 since the majority of cases occured for TRAUMATIC INJURIES AND DISORDERS with 8535 cases.

```
cons_nature <- cons_v3 %>%
  select(ID, Final_Narrative, nature_L2) %>%
  mutate( Nature = case_when(
   nature_L2 == "Surface wounds and bruises" ~ "Open wounds",
   nature_L2 == "Traumatic injuries and disorders, unspecified" ~ "Other traumatic injuries and disorde
    nature_L2 == "Circulatory system diseases" ~ "Other - Nature",
   nature_L2 == "Respiratory system diseases" ~ "Other - Nature",
   nature_L2 == "Disorders of the skin and subcutaneous tissue" ~ "Other - Nature",
   nature_L2 == "Symptoms" ~ "Other - Nature",
   nature_L2 == "EXPOSURES TO DISEASE-NO ILLNESS INCURRED" ~ "Other - Nature",
   nature_L2 == "Bacterial diseases" ~ "Other - Nature",
   TRUE ~ as.character(nature_L2)))

cons_nature %>%
  count(Nature, sort = T)
```

Nature	n
Traumatic injuries to bones, nerves, spinal cord	3117
Open wounds	2166
Other traumatic injuries and disorders	1662
Burns and corrosions	533
Intracranial injuries	325
Multiple traumatic injuries and disorders	318
Effects of environmental conditions	288
Traumatic injuries to muscles, tendons, ligaments, joints, etc.	126
Other - Nature	28

```
# since the count is small, we'll remove "Other - Nature"
cons_nature_2 <- cons_nature %>%
  select(-nature_L2) %>%
  filter (Nature!= "Other - Nature")

# write.csv(cons_nature_2, "Final_Prepetation/cons_nature_2.csv", na = "")
```