Investigating Data Engineering, Feature Extraction and Machine Learning methods to code MedDRA dictionaries

29NOV2019

Abstract and Understanding MedDRA

INTRODUCTION

- Since the 1990's MedDRA dictionaries have been used to code standard terminology.
- With advancement in technologies and the growing need for better world wide health care, can machine learning and data manipulation methods be used to code MedDRA dictionaries?
- Machine learning models were created based on features extracted from phrases commonly used in three different languages, namely English, Afrikaans and Dutch.
- With high accuracy the model is able to predict to which language the phrase belongs.
- A similar approach can be applied to MedDRA coding.

Abstract and Understanding MedDRA

MEDDRA BASICS

- MedDRA dictionaries consist of 5 logically structured levels, arranged from general to specific terminology available in several languages.

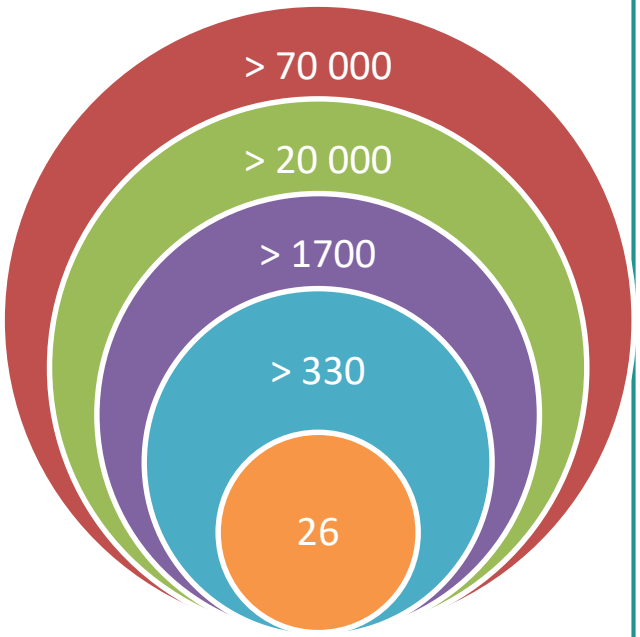
Lowest Level Term (LLT) – Assigns MedDRA codes to reported observations at database level.

Preferred Term (PT) - single concept characteristics such as symptoms, signs, diagnosis, indications, investigations, etc.

High Level Term (HLT) – based on anatomy, pathology, physiology, etiology or function .

High Level Group Term (HLGT) – grouped into SOC's.

System Organ Class (SOC) – grouped by etiology, site or purpose.

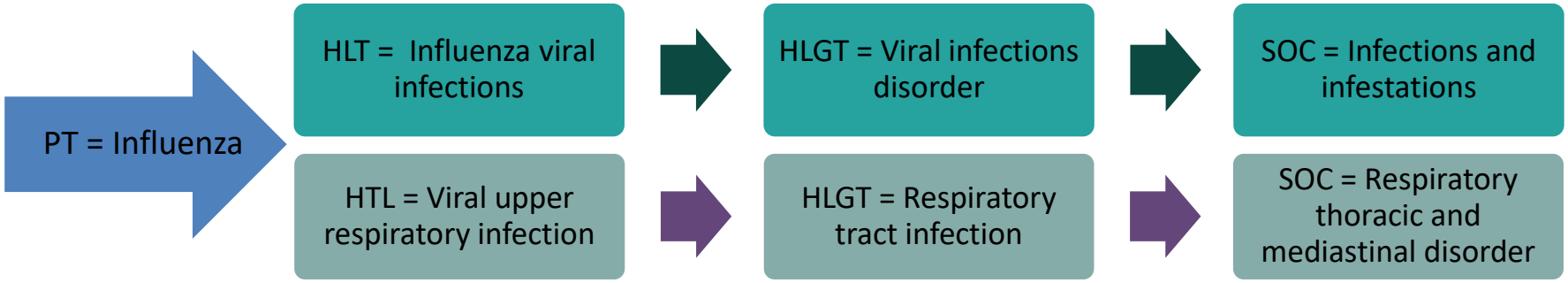


Abstract and Understanding MedDRA – cont.

ONE OF THE MEDDRA CHALLENGES

- The MedDRA structure is “multiaxial”, meaning the preferred term “Influenza” can represent a “viral infection” as well as a “ respiratory infection”.

“For this reason, each PT is assigned to a primary SOC, but may also be assigned to one or more secondary SOCs.” – extract from Understanding medDRA – www.meddra.org



Source: Understanding MedDRA – www.meddra.org

- Instances of the above example may occur often, it is therefore important to keep this in mind when building the ML model. Classifying multiple outcomes from a single input will be one of many challenges.

Language Dataset

THE LANGUAGE DATASET

- The language dataset consist of common phrases in English, Afrikaans and Dutch. In total the dataset consists of 2671 phrases.

- There are 2055(74.43%) English phrases, 639(23.14%) Afrikaans phrases and 67(2.43%) Dutch phrases.

- Why use the language dataset ?
  - The datasets is riddled with common errors such as spelling mistakes, incomplete words, special characters and this equates to real world data
  - Afrikaans and Dutch is closely related, making classification tricky
  - Unequal amount of phrases per language skewing the dataset

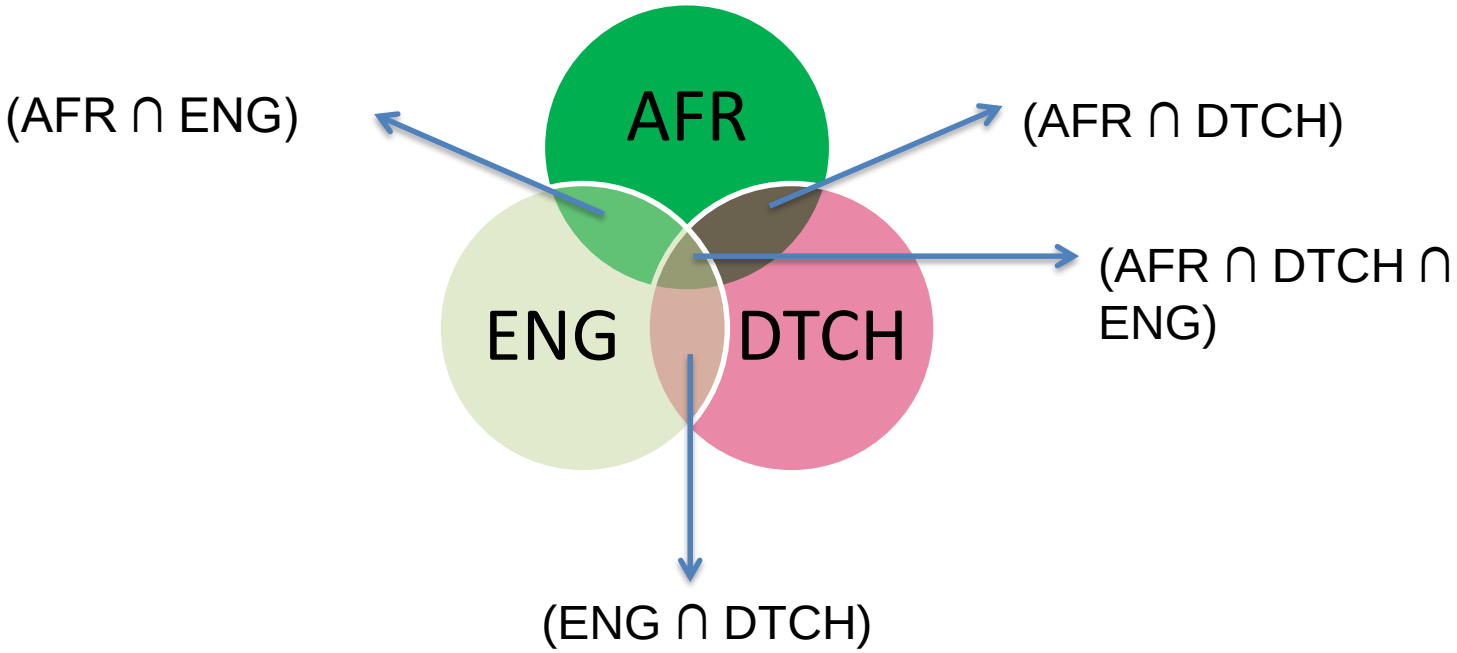
(A) bigger bang for your buck	English
Ek sal kan doen met â€˜n miljoen.	Afrikaans
Eigen haard is goud waard.	Dutch
Ek is in die dolliwarie.	Afrikaans
Een ongeluk komt zelden alleen.	Dutch
A diamond in the rough	English

Snapshot of original data

Language Dataset - continued

THE LANGUAGE DATASET continued

- Main idea and focus points while engineering and extracting features from the data.



DATA CLEANING, ENGINEERING and FEATURE EXTRACTION

DATA CLEANING

- To achieve optimal model accuracy the data were cleaned and standardized. All data were uppercased, spaces and special characters were removed.
- Lucky for us MedDRA dictionaries consist of clean data, but unfortunately spelling, accuracy and exactness is not a common human trait even with best intentions.

DATA ENGINEERING and FEATURE EXTRACTION

Below are two examples of data engineering and extraction :

- sum_lett = Sum of the letters where each letter is assigned a numerical value (A=1,B=2,…,Z=26). The logic behind this variable is simple. For example “PHUSE” can be spelled in 120 ways, one of which is correct but the remaining 119 options are incorrect. If we assume all of the correct letters were used in spelling “PHUSE” then PHUSE=PUHSE=…=ESUHP=69. This is an example of data engineering.
- num_wrds = Number of word in each phrase. This is an example of feature extraction.

DATA CLEANING, ENGINEERING and FEATURE EXTRACTION - continued

LANGUAGE DICTIONARIES

- For each language a word dictionary was created. The dictionaries consists of a record for each and every unique word present in the phrase by language.
- The below set of variables were created in conjunction with the language dictionaries.
  - num_eng = Number of English words in the phrase.
  - num_afr = Number of Afrikaans words in the phrase.
  - num_dtch = Number of Dutch words in the phrase.

Is this data engineering, feature extraction or both?

text	language	Text_comp	sum_lett	num_wrds	num_eng	num_afr	num_dtch
A bigger bang for your buck	English	ABIGGERBANGFORYOURBUCK	228	6	6	2	1
Ek sal kan doen met n miljoen	Afrikaans	EKSALKANDOENMETNMILJOEN	242	7	1	7	2
Eigen haard is goud waard	Dutch	EIGENHAARDISGOUDWAARD	194	5	1	2	5
Ek is in die dolliwarie	Afrikaans	EKISINDIEDOLLIWARIE	193	5	3	5	2
Een ongeluk komt zelden alleen	Dutch	EENONGELUKKOMTZELDENALLEEN	283	5	0	3	5
A diamond in the rough	English	ADIAMONDINTHEROUGH	186	5	5	2	2

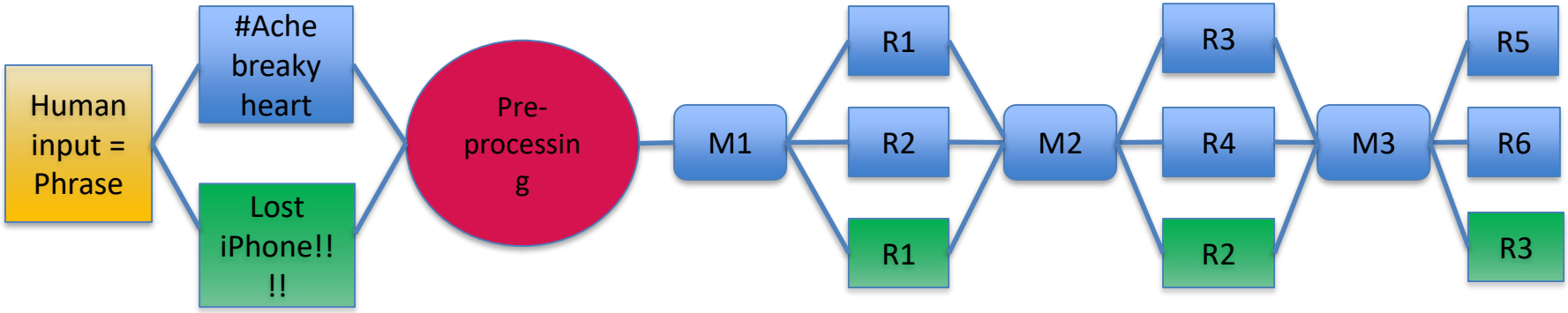
Snapshot of some data used in the model

MACHINE LEARNING (ML) APPROACH AND MODEL

LANGUAGE APPROACH IN TERMS OF MEDDRA CLASSIFICATION

- How is the MedDRA classification tied to the Language phrase classification? First we have human input but thereafter the model classification becomes our new human input. This leads to a multilayered model approach when classifying MedDRA terms.

*** Mx = Model x; Rx = Possible outcome x for each user input**



- The goal is therefore to classify the human input with one or several LLT's, then HLT's, then HLGT's and finally SOC's. For the language dataset the phrase is classified to the outcome with the highest probability but for MedDRA classification, human input will have to be paired with several possible outcomes based on a probability threshold.

MACHINE LEARNING (ML) APPROACH AND MODEL - continued

THE MODEL

- Several models were fitted with each approach displaying distinct advantages and disadvantages. The final approach was a combination of various models to produce the best overall accuracy and speed when classifying the phrase.

- The accuracy of the initial model was measured by employing Leave-One-Out Cross-Validation (LOOCV). The CV-error of the model had an error rate of 2.5%, which equates to 69 misclassification from 2671 phrases. That is 97.5% of the phrases were classified correctly.

- To further improve on the above approach some variables were transformed, interactions terms were added and re-sampling of the initial data was introduced. These few changes in the model dramatically improved the models accuracy. The final model had a CV-error of 0.0007%, that is only 2 misclassifications with an accuracy of 99.3%.

MACHINE LEARNING (ML) APPROACH AND MODEL - continued

THE MODEL continued

text	language	sum_lett	num_wrds	num_eng	num_afr	num_dtch	prediction	afr_prob	dtch_prob	eng_prob
Youll wonder where the yellow went when you brush your teeth with Pepsodent	English	900	13	13	1	0	Afrikaans	0.547855	0	0.452145
Some men are born mediocre some men achieve mediocrity and some men have mediocrity thrust upon them	English	965	17	17	3	3	Afrikaans	0.999108	0	0.000892

Snapshot of the misclassifications based on LOOCV

- However, be aware, the above model's accuracy can be misleading…
- In the testing phase it was found the model will classify to the majority language, in this case to "English" if a new phrase not present in the current language dataset is entered. A similar issue will arise with MedDRA classification, if new input not present in the current data is entered. **So a new challenge has made it's appearance…and that is the beauty of Machine learning, you learn as much as the model if not more!**

MACHINE LEARNING (ML) APPROACH AND MODEL - continued

WHERE TO IMPROVE

- Some literature suggest alternative approaches to the one followed in the presentation.
  - For example: naïve Bayes and Support vector models
  - Combine methods and do not rely on only one approach.
- Consider other re-sampling methods for skewed data.
- Get more data, especially if the data is skewed.
- Cryptography.
- Geographic user information.
- Reinforced learning.
- How to handle single word input.

MACHINE LEARNING (ML) APPROACH AND MODEL - continued

MODEL WITHOUT RE-SAMPLING

Afrikaans phrase:

- Coenraad Bezuidenhout

```
> new()
Enter phrase: Coenraad Bezuidenhout
[1] "The phrase is classified as : English (1.6e-05, 0, 0.999984)"
Is the classification correct? (Y=Yes or N=No): N
Please enter the correct language of the phrase (1=English, 2=Afrikaans, 3=Dutch): 2
[1] "Adding new phrase to language dictionaries and dataset"
```

- Coen Bezuidenhout

```
> new()
Enter phrase: Coen Bezuidenhout
[1] "The phrase is classified as : English (0.18528, 0, 0.81472)"
Is the classification correct? (Y=Yes or N=No): N
Please enter the correct language of the phrase (1=English, 2=Afrikaans, 3=Dutch): 2
```

MACHINE LEARNING (ML) APPROACH AND MODEL - continued

MODEL WITH RE-SAMPLING

Afrikaans phrase:

- Coenraad Bezuidenhout

```
> new_smote()
Enter phrase: Coenraad Bezuidenhout
[1] "The phrase is classified as : English (0.001043, 0, 0.998957)"
Is the classification correct? (Y=Yes or N=No): N
Please enter the correct language of the phrase (1=English, 2=Afrikaans, 3=Dutch): 2
```

- Coen Bezuidenhout

```
> new_smote()
Enter phrase: Coen Bezuidenhout
[1] "The phrase is classified as : Afrikaans (0.911019, 0, 0.088981)"
```

MACHINE LEARNING (ML) APPROACH AND MODEL - continued

MODEL WITHOUT RE-SAMPLING

Dutch phrase:

- Een koffie alstublieft

```
> new()
Enter phrase: Een koffie alstublieft
[1] "The phrase is classified as : Afrikaans (0.998188, 0, 0.001812)"
Is the classification correct? (Y=Yes or N=No): N
Please enter the correct language of the phrase (1=English, 2=Afrikaans, 3=Dutch): 3
```

- koffie alstublieft

```
Enter phrase: koffie alstublieft
[1] "The phrase is classified as : Dutch (3e-06, 0.991802, 0.008195)"
Is the classification correct? (Y=Yes or N=No): Y
```

- alstublieft

```
Enter phrase: alstublieft
[1] "The phrase is classified as : English (0, 6.3e-05, 0.999937)"
Is the classification correct? (Y=Yes or N=No): N
```

MACHINE LEARNING (ML) APPROACH AND MODEL - continued

MODEL WITH RE-SAMPLING

Dutch phrase:

- Een koffie alstublieft

```
> new_smote()
Enter phrase: Een koffie alstublieft
[1] "The phrase is classified as : Afrikaans (0.999883, 4e-06, 0.000114)"
Is the classification correct? (Y=Yes or N=No): N
Please enter the correct language of the phrase (1=English, 2=Afrikaans, 3=Dutch): 3
```

- koffie alstublieft

```
> new_smote()
Enter phrase: koffie alstublieft
[1] "The phrase is classified as : Dutch (0, 0.999995, 5e-06)"
Is the classification correct? (Y=Yes or N=No): Y
```

- alstublieft

```
> new_smote()
Enter phrase: alstublieft
[1] "The phrase is classified as : English (0, 0.186342, 0.813658)"
Is the classification correct? (Y=Yes or N=No): N
Please enter the correct language of the phrase (1=English, 2=Afrikaans, 3=Dutch): 3
```

QUESTIONS