Project progress

Contents

[Feb 06 2018 1](#_Toc505700385)

[Examples of incorrect values: 2](#_Toc505700386)

## Feb 06 2018

A set of patients was provided in [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_1]

|  |
| --- |
| select count(distinct patientsid) PatientCount  from [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_1]  -- 7290 |

Split patients into 4 categories:

* 1500 patients in train set SetID = 1
* 1500 patients in dev set SetID = 2
* 1500 patients in test set SetID = 3
* 2653 patients in other set SetID = 4

137 patients did not have TIUDocuments in the targed date range.

Pulled all TIUDocuments in time range

TimeWindowStart - 2 > ReferenceDateTime >TimeWindowEnd + 2. If TimeWindowEnd was null, it was assumed Feb 1, 2018.

Total doc count = 1,091,009. Out of those, 1,032,889 TIUDocuments had ReportText not null.

Into [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_1\_PtDoc\_Sets]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SetID | Min\_RowIndex | Max\_RowIndex | PateintCount | DocumentCount |
| 1 | 1 | 211969 | 1500 | 211969 |
| 2 | 211970 | 414325 | 1500 | 202356 |
| 3 | 414326 | 614976 | 1500 | 200651 |
| 4 | 614977 | 1032889 | 2653 | 417913 |

|  |
| --- |
| select count(distinct TIUDocumentsid), count(distinct patientsid)  FROM [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_1\_PtDoc\_Sets]  TIUDoc count = 1032889 Patient count = 7153 |

Pulled documents for all patients in the enriched set limiting the document window to

TimeWindowStart -1… TimeWindowEnd +1 days

|  |
| --- |
| SELECT [PatientSID]  ,[PatientIEN]  ,[Sta3n]  ,[BirthDateTime]  ,[DeceasedFlag]  ,[DeathDateTime]  ,[Gender]  ,case when [TimeWindowStart] is null then '2000-01-01' else [TimeWindowStart] end [TimeWindowStart] -- No null was found  ,case when [TimeWindowEnd] is null then '2018-02-01' else [TimeWindowEnd] end [TimeWindowEnd] -- Some records have null  ,[StrokeYN]  FROM [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_1] ) c  join [CDWWork].[TIU].[TIUDocument\_8925] t on c.PatientSID=t.PatientSID  where datediff(day, [TimeWindowStart], ReferenceDateTime ) > -2  and datediff(day, ReferenceDateTime, [TimeWindowEnd] ) > -2 |

Processed all documents in Set 1.

|  |
| --- |
| SELECT [TIUDocumentSID]  ,[AnnotationType]  ,[anchor]  ,[InstanceID]  ,[Snippets]  ,[SpanStart]  ,[SpanEnd]  FROM [MAVIN\_ComputeLib].[NLP\_Stroke].[ScoreOut\_20180201] |

N=3,898 (4,460 with duplicates) scores were found in 3562 documents (out of 211,969 in set 1) for 920 (out of 1500) patients.

Precision validation was performed on a random set of 300 documents that had score values. Out of 383 reviewed instances, 321(83.8%) instances were found to be correct, 24 (6.3%) were found to be related to NIH score but with incorrectly pulled numerical value, and 38 (9.9%) were found to be fully unrelated to NIH stroke score.

### Examples of incorrect values:

In the sentence below, NIHSS score was extracted as 2 and not at 14.

|  |  |
| --- | --- |
| **Statement** | **Extracted value for NIHSS** |
| NIHSS 14 for VF 2, Face 2, RUE 1, RLE 0, LUE 3, LLE 2, Sens 1, Dysarthria 1, Extinction | 2 |
| NEUROLOGICAL: NIHSS: 18 1 (Partial gaze palsy) | 1 |
| NIHSS level of consciousness 0 ask month and age 1 (missed month) Blink eye/squeeze:0 horizontal extraocular movements 1 visual fields 1 facial palsy 2 left arm drift 4 right arm motor drift 0 left leg motor drift 1 right leg motor drift 0 limb ataxia 0 sensation 0 language 0 dysarthria 1 extinction 1 | 0 (no total was listed in the document) |
| NIHSS: 2 (1-L partial hemianopia, 1-dysarthria). | 1 |
| NIH 2: Sensory 1, ataxia 1 | 1 |

Examples of irrelevant values:

|  |
| --- |
| Diagnosis:Stroke ICD-10 |
| Subacute R MCA stroke LDL 85 (at UAMS) |
| Stroke 1 called, neurology received page at 1pm |
| Multiple Strokes (4 -> most recent xx/xx/xx) |

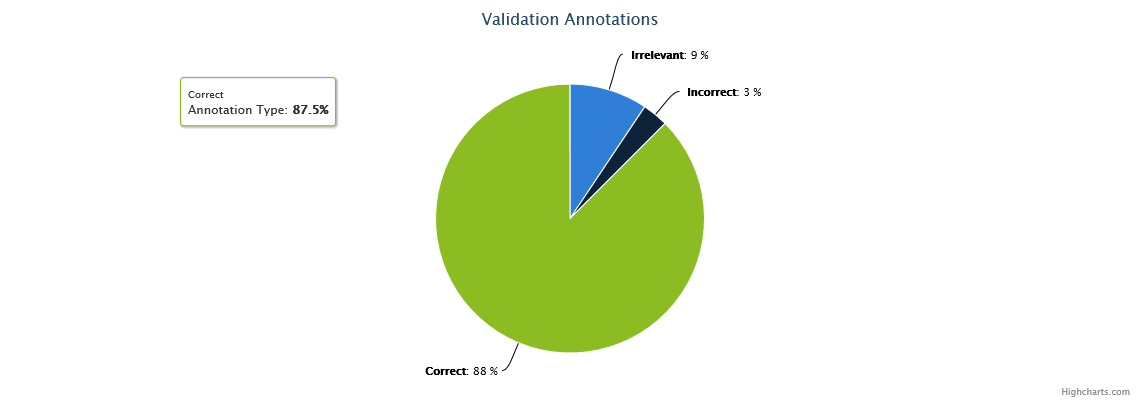
Mar 4

Iteration 2

**Total Validations: 96**

Validation Annotations

Irrelevant: 9 %Incorrect: 3 %Correct: 88 %CorrectAnnotation Type: 87.5%Highcharts.com



| Value | Count | % |
| --- | --- | --- |
|  |  |  |
| [Correct](http://vhacdwdwhweb11:8080/chex/reports/validationDetail?validationValue=Correct&report=true) | 84 | 87.5 |
| [Incorrect](http://vhacdwdwhweb11:8080/chex/reports/validationDetail?validationValue=Incorrect&report=true) | 3 | 3.16 |
| [Irrelevant](http://vhacdwdwhweb11:8080/chex/reports/validationDetail?validationValue=Irrelevant&report=true) | 9 | 9.47 |

SELECT identity(int, 1,1) ScoreID, NIHSS\_score\_value, Snippets, ReferenceDateTime, PatientSID, Sta3n, a.TIUDocumentSID

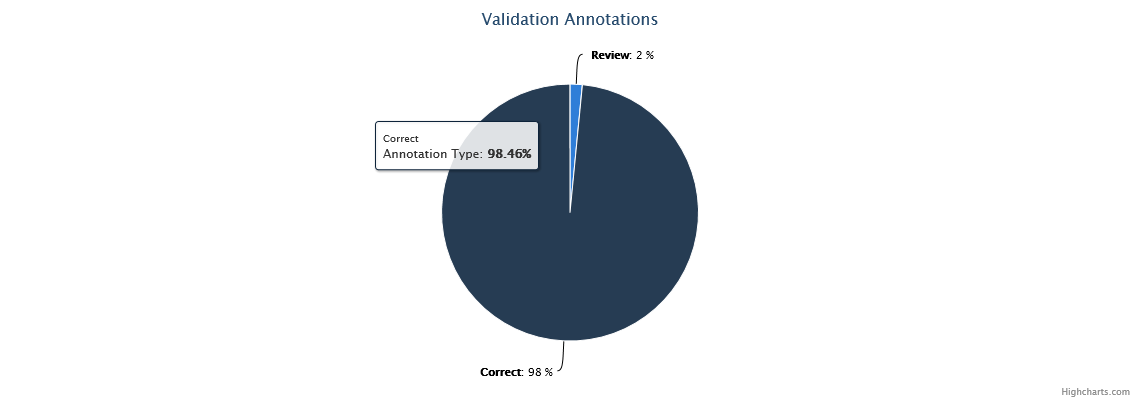
into [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_1\_setid1\_StrokeScore\_20180228]

FROM (select distinct TIUDocumentSID, NIHSS\_score\_value, Snippets from [MAVIN\_ComputeLib].[NLP\_Stroke].[ScoreOut\_Set1\_20180228] ) a

join [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_1\_PtDoc\_Sets] b

on a.TIUDocumentSID=b.tiudocumentsid

Validation 3 was completed after fixing issues identified by validation2



Validation 4.

200 documents were selected for recall validation. 100 docs from set 1 and 100 docs from set 2.

|  |
| --- |
| SELECT distinct \*  --into [MAVIN\_ComputeLib].[NLP\_validation].[enrichedPtSet\_1\_2\_ValidationSetid3]  from (  select top 100 [PatientSID],[TIUDocumentSID], reporttext,[ReferenceDateTime]  from ( select a.\*, reporttext  FROM [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_1\_PtDoc\_Sets] a  join [cdw\_tiu].[TIU].[TIUDocument\_8925\_02] b with(nolock) on a.tiudocumentsid=b.tiudocumentsid  where setid=3 and (contains(reporttext, 'nih\*') or contains(reporttext, '"national institute\*"'))) n order by newid()  union  select top 100 [PatientSID],[TIUDocumentSID], reporttext ,[ReferenceDateTime]  from ( select a.\*, reporttext  FROM [MAVIN\_ComputeLib].[NLP\_Stroke].[enrichedPtSet\_2\_PtDoc\_Sets] a  join [cdw\_tiu].[TIU].[TIUDocument\_8925\_02] b with(nolock) on a.tiudocumentsid=b.tiudocumentsid  where setid=3 and (contains(reporttext, 'nih\*') or contains(reporttext, '"national institute\*"')) ) m order by newid() ) k |

Results

Out of 200 documents,

NLP found 86 docs/106 instances ( 49 docs/60 instances in set 1, 41 docs/50 instances in set 2)

RefStandard had 153 documents/201 instances (88 docs/116 instances from Set1, 74 docs/94 instances in Set2).

4 documents are in both sets.

Instance level efficacy:

LTP:98 LFP:8 LFN:103 LPrecision: 0.9245 LRecall: 0.4876 LF1 0.6384