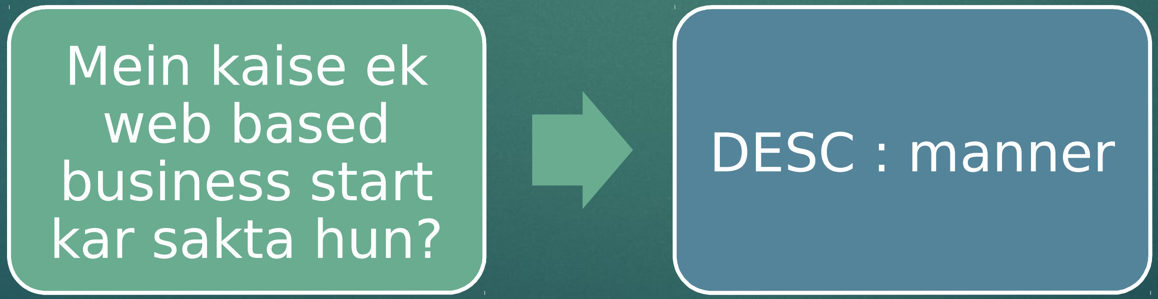
Question classification for Code-Mixed text

ARCHIT RATHORE

PRABUDDHA CHAKRABORTY

Problem Statement

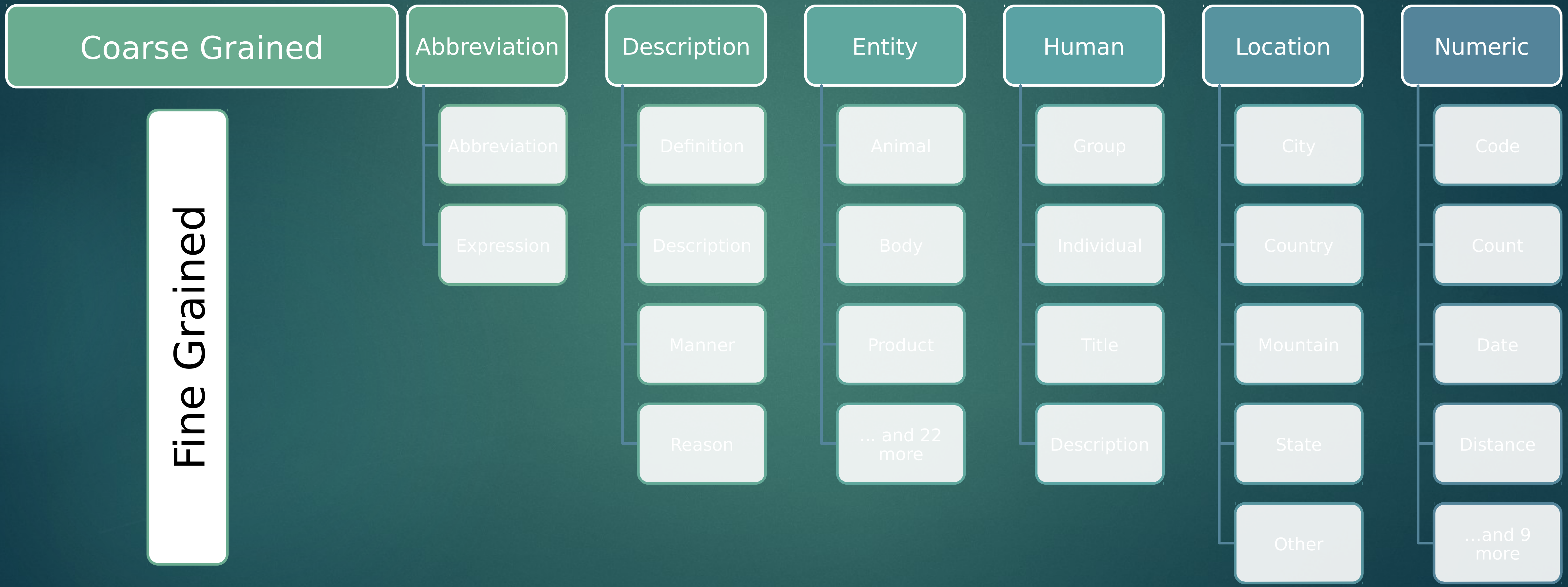
- Given:
  - A code-mixed question
- To predict:
  - Expected type of the answer



Mein kaise ek web based business start kar sakta hun?

DESC : manner

Answer-type hierarchy [1]



Coarse Grained

Fine Grained

- Abbreviation
  - Abbreviation
  - Expression
- Description
  - Definition
  - Description
  - Manner
  - Reason
- Entity
  - Animal
  - Body
  - Product
  - … and 22 more
- Human
  - Group
  - Individual
  - Title
  - Description
- Location
  - City
  - Country
  - Mountain
  - State
  - Other
- Numeric
  - Code
  - Count
  - Date
  - Distance
  - …and 9 more

Challenges

- No dataset available on the public domain
- Dearth of tools that handles code-mixed text – no POS-taggers, chunkers, dependency parsers, language identification tools etc

Proposed Solution

- Create a new dataset of code-mixed questions
- Make dataset in a format compatible with the existing tools

Dataset Format

- QueNo#1 CoarseLabel:FineLabel <QueString in English>
- QueNo#2 CoarseLabel:FineLabel <QueString in Hindi>
- QueNo#3 CoarseLabel:FineLabel <QueString in Codemixed_scriptPreserved>
- QueNo#4 CoarseLabel:FineLabel <QueString in Codemixed_romanized>

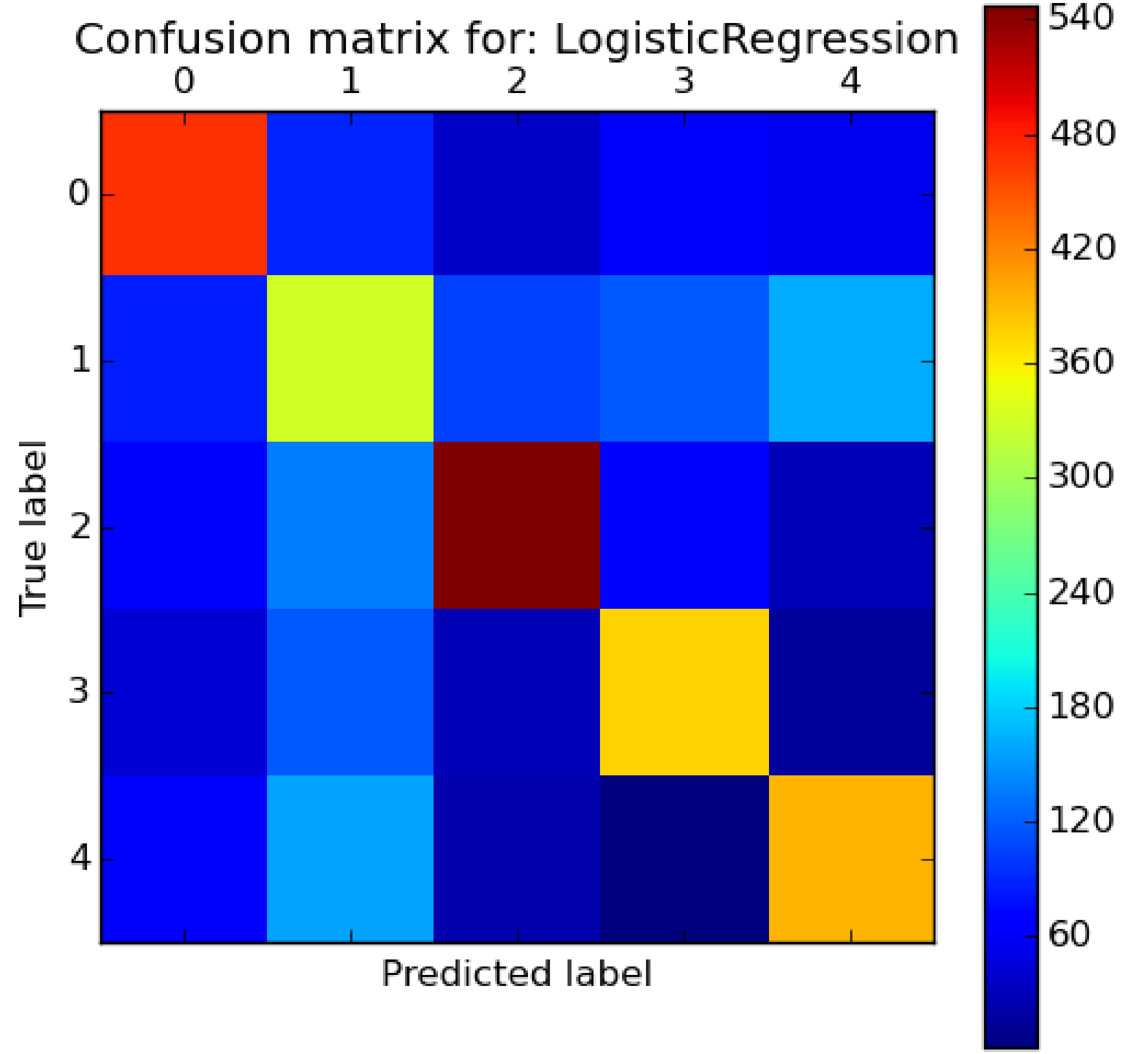
- 185#1 DESC:reason Why do horseshoes bring luck ?
- 185#2 DESC:reason घोड़े की नाल भाग्य क्यों लाती है?
- 185#3 DESC:reason घोड़े की नाल luck क्यों लाती है?
- 185#4 DESC:reason Ghode ki naal luck kyon laati hai?

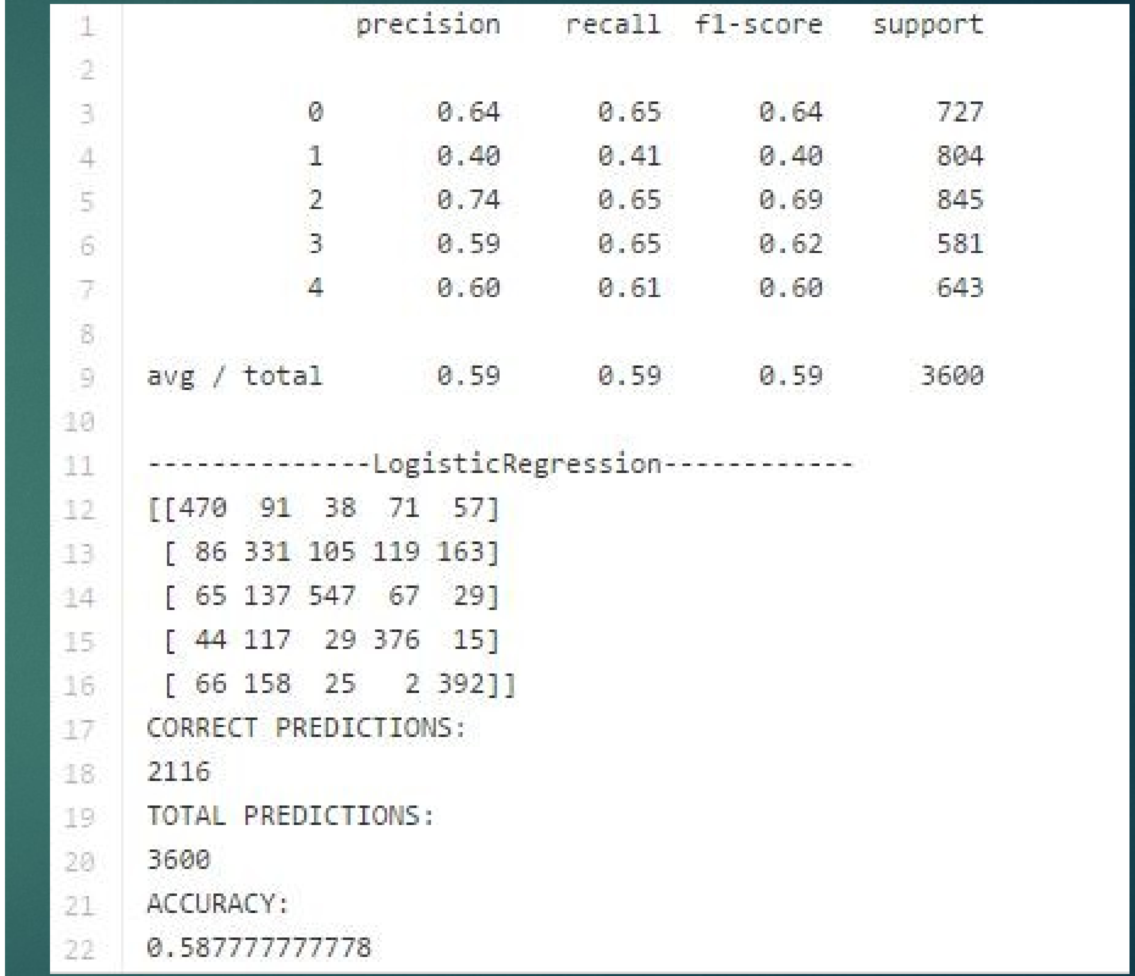
Dataset Creation

- Get English annotated questions from dataset created by Li and Roth [5]
- Use python’s Goslate API to convert these questions to Hindi
- Manually do the following:
  - Fix translation errors in Goslate output
  - Repose the questions in a code mixed sense
- Transliterate the script-preserved code-mixed questions to roman script
  - Done using “Sanscript” API : has problems with schwa deletion [8]
  - Manually corrected as of now

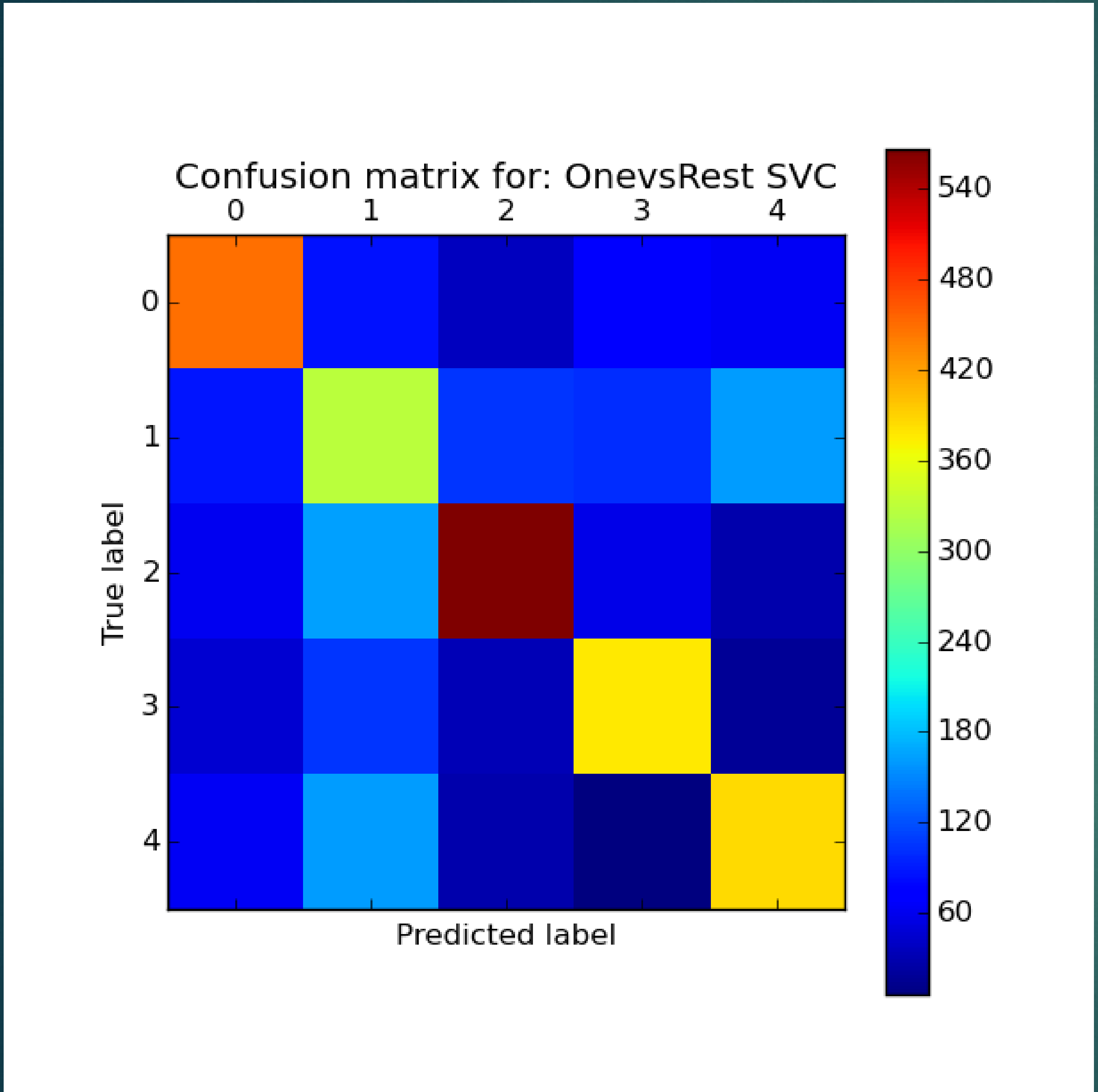
Preliminary Results

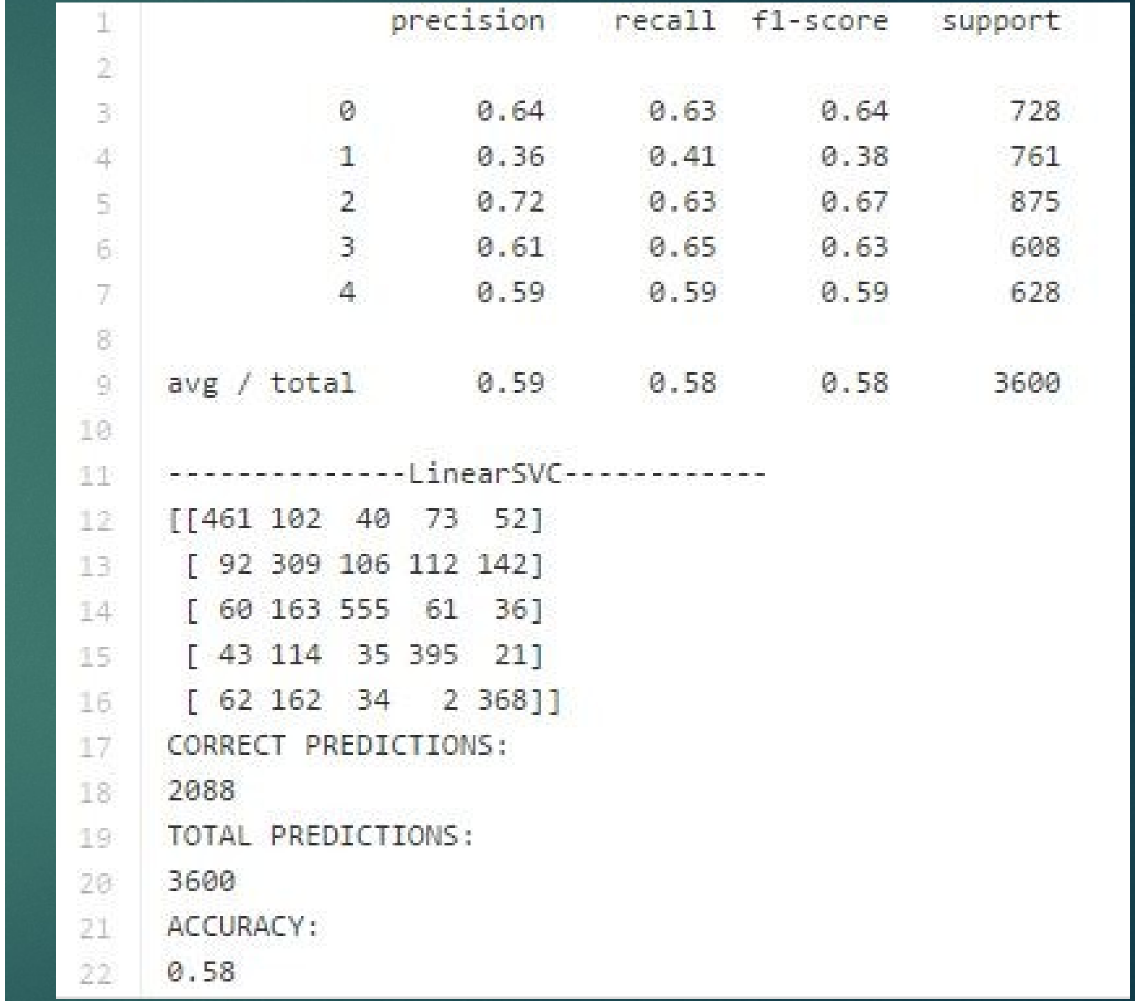
- Data of 200 code-mixed sentences
- Used :
    - Linear kernel SVM   (one-vs-rest approach)
    - Logistic regression classifier
- Performed 100 iterations of training and testing with both classifiers
- Split data randomly with 0.8 : 0.2 ratio





```
             precision    recall  f1-score   support

          0       0.64      0.65      0.64       727
          1       0.40      0.41      0.40       804
          2       0.74      0.65      0.69       845
          3       0.59      0.65      0.62       581
          4       0.60      0.61      0.60       643

avg / total       0.59      0.59      0.59      3600

--------------LogisticRegression-----------
[[470  91  38  71  57]
 [ 86 331 105 119 163]
 [ 65 137 547  67  29]
 [ 44 117  29 376  15]
 [ 66 158  25   2 392]]
CORRECT PREDICTIONS:
2116
TOTAL PREDICTIONS:
3600
ACCURACY:
0.587777777778
```

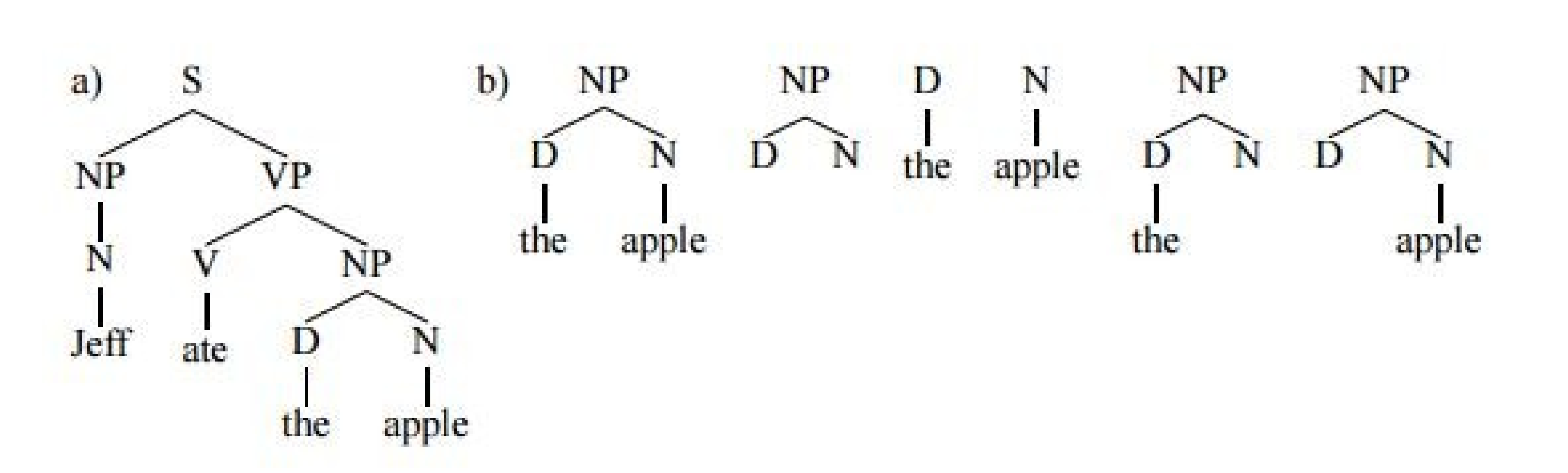




```
             precision    recall  f1-score   support

          0       0.64      0.63      0.64       728
          1       0.36      0.41      0.38       761
          2       0.72      0.63      0.67       875
          3       0.61      0.65      0.63       608
          4       0.59      0.59      0.59       628

avg / total       0.59      0.58      0.58      3600

--------------LinearSVC-----------
[[461 102  40  73  52]
 [ 92 309 106 112 142]
 [ 60 163 555  61  36]
 [ 43 114  35 395  21]
 [ 62 162  34   2 368]]
CORRECT PREDICTIONS:
2088
TOTAL PREDICTIONS:
3600
ACCURACY:
0.58
```

Further Work

- Use a tree kernel for SVM as proposed by Collin and Duffy [3]



- Introduce adjacency features as proposed by Raghavi et al [2]
- Use DCNN for sentence modelling – Kalbrechhner et al [6]

Fig – Collin and Duffy

References

1. Xin Li and Dan Roth. “Learning question classifiers”. Proceedings of the 19th international conference on Computational linguistics, 2002.
2. Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. “Answer ka type kya he?” Learning to classify questions in code-mixed language. Proceedings of the 24th International Conference on World Wide Web, pages 853–858, 2015
3. Collins, Michael, and Nigel Duffy. "Convolution kernels for natural language.“ *Advances in neural information processing systems*. 2001.
4. Siva Reddy. Hindi dependency parser. https://bitbucket.org/sivareddyg/ hindi-dependency-parser, 2014
5. Xin Li and Dan Roth. Experimental data for question classification. http://cogcomp.cs.illinois.edu/Data/QA/QC/.
6. Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." *arXiv preprint arXiv:1404.2188* (2014).
7. Source - https://github.com/sanskrit/sanscript