

Ensemble Classifier based Approach for Code-Mixed Cross-Script Question Classification

Team : IINTU

Debjoyti Bhattacharjee

School of Computer Science and
Engineering
Nanyang Technological University
Singapore



**NANYANG
TECHNOLOGICAL
UNIVERSITY**

Paheli Bhattacharya

Department of Computer Science and
Engineering
Indian Institute of Technology Kharagpur
India



OUTLINE OF THE PRESENTATION

- Mixed Script Information Retrieval (MSIR)
- Question Classification in Code-Mixed data
- Proposed Approach
- Experimental Setup
- Results
- Conclusion and Future Work

MIXED-SCRIPT/ CODE-MIXED DATA

ADJUST KIJIYE
EK CHANCE MILEGA?



BAHUT TENSION HAI

"Bahut Garmi Hai, Yaar!"

MIXED-SCRIPT/CODE-MIXED DATA

- Both documents and queries are in more than one scripts
- **Transliterated** from native script (Devnagari for Hindi) to foreign script (Roman)
- Define MSIR formally ¹ :
 - Natural languages $\mathbf{L} = \{l_1, l_2, \dots, l_n\}$
 - Scripts $\mathbf{S} = \{s_1, s_2, \dots, s_n\}$
such that s_i is the native script for language l_i
 - Word $w_i = \langle l_i, s_j \rangle$
 - $i = j$, **native script** , else transliterated

¹Gupta et. al. , Query Expansion for Mixed-Script Information Retrieval, SIGIR 2014

WHY MSIR ?

- Users now opt to write in their native language rather than English
- Shortcoming : Font-encoding issues, English keyboard
- Write in the Roman Script by transliteration

QUESTION CLASSIFICATION

- Question Answering
 - Find concise and accurate answer to a given question
- Question Classification
 - Subtask of Question Answering
 - Determine the type of answer for a question
- Categorize a question in to a set of classes and deal with each class for answering

CODE-MIXED CROSS-SCRIPT QUESTION CLASSIFICATION

- Mixing of the languages English and Bengali
- Set of questions $Q = \{q_1, q_2, \dots, q_n\}$
- Each question $q = \langle w_1 w_2 \dots w_n \rangle$
 - w_i = English word or transliterated Bengali
- Set of classes $C = \{c_1, c_2, \dots, c_m\}$
- Classify question q_i to a class c_j

QUESTION CLASSIFICATION IN MIXED-SCRIPT

Kharagpur theke Howrah car fare koto?

Bengali

English

DISTANCE

TEMPORAL

MONEY

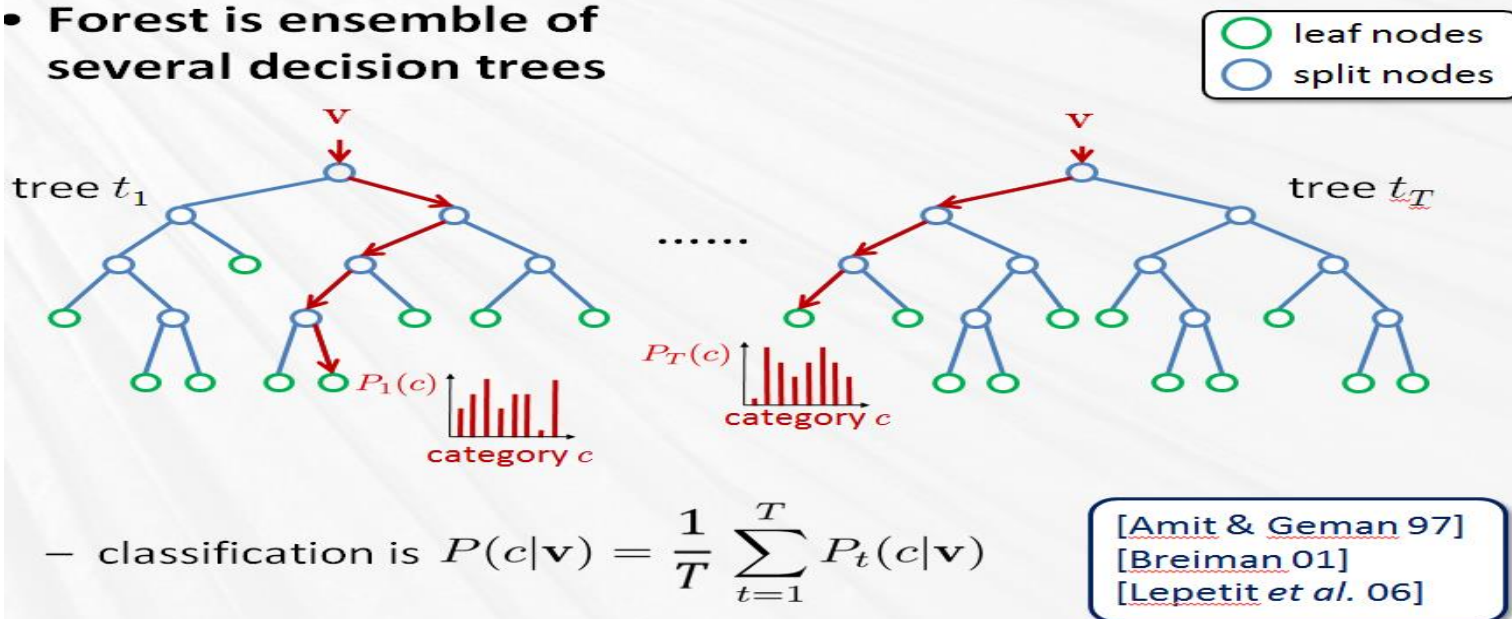
LOCATION

PROPOSED APPROACH

- Each question is represented as a 2000 dimensional binary vector
 - i^{th} component \Leftrightarrow the i^{th} most frequent word
- Train classifiers
 - Random Forests (RF)
 - One-Vs-Rest (OvR)
 - k-Nearest Neighbours (kNN)
- Ensemble of the classifiers
 - Majority Vote
 - Else, a random label
- Retraining
 - From the test set, pick up 90% of the samples (by replacement) which had the same label for all the 4 classifiers
 - New training = Original Training Set + Sampled Test Set

RANDOM FOREST (RF)

- Forest is ensemble of several decision trees






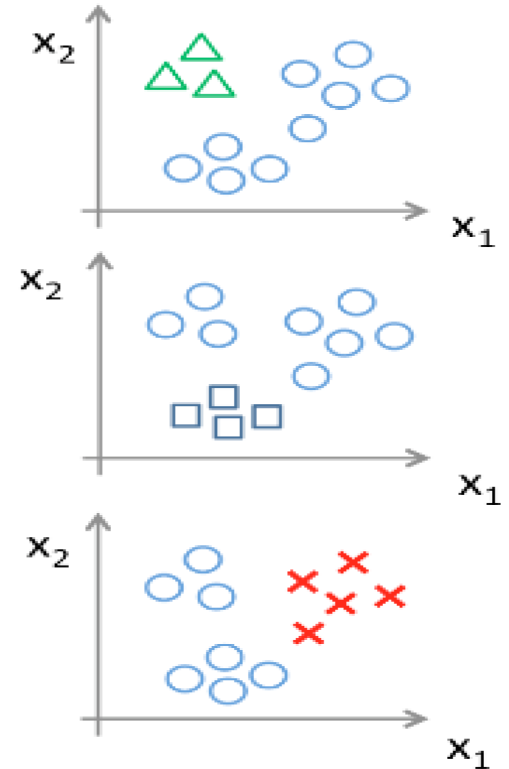
- Ensemble learning method
- Fits a number of decision trees on various sub-samples of the dataset
- Use averaging to improve the predictive accuracy and control overfitting

ONE-VS-REST (OVR)

One-vs-all (one-vs-rest):

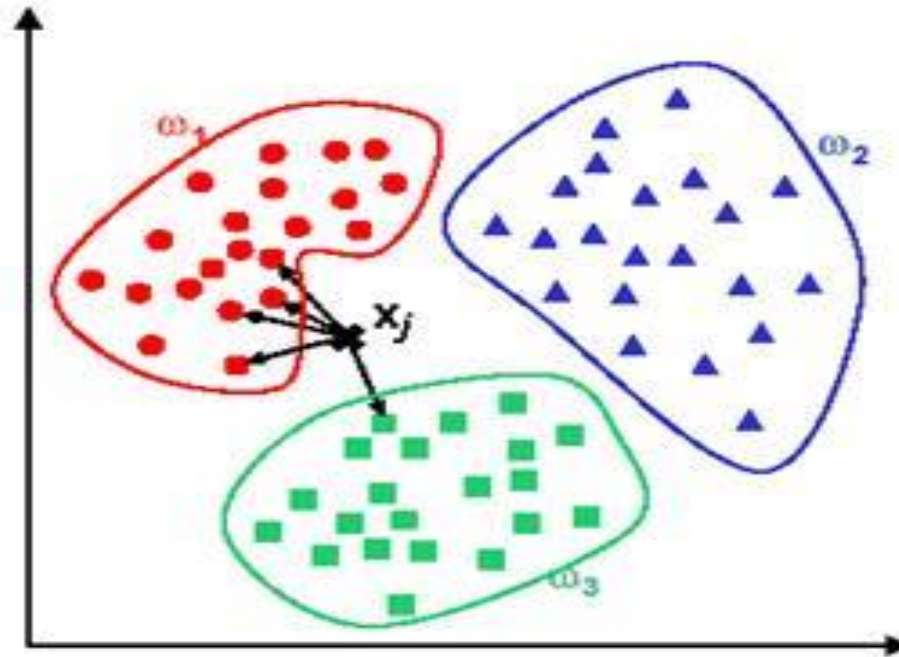


Class 1: 
Class 2: 
Class 3: 



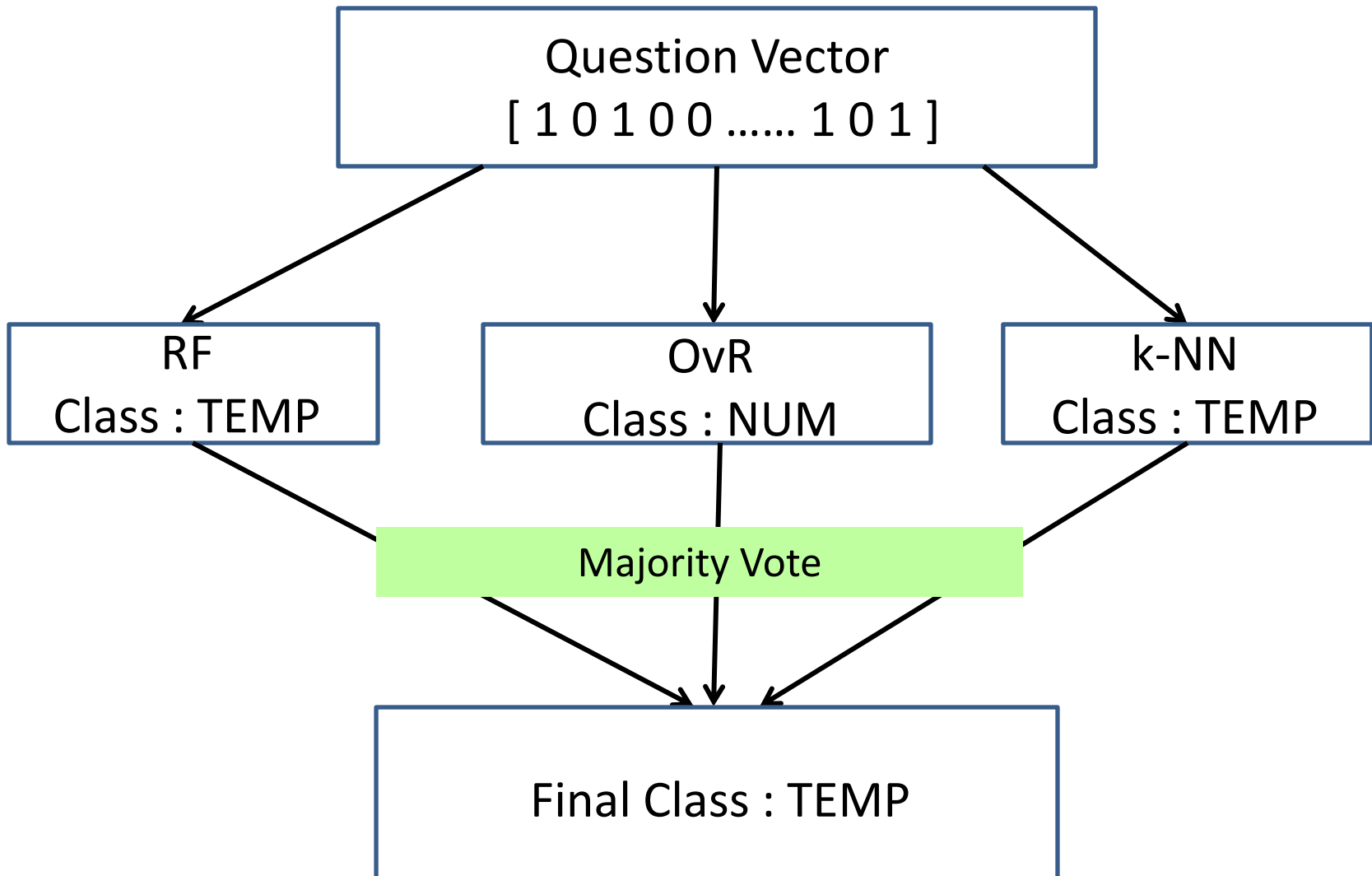
- Fits one classifier per class i to predict $p(\text{class}=i \mid x, \theta)$
- Test sample, pick the class i that has the maximum probability
- Each classifier is trained with the entire dataset
- Most commonly used strategy for multiclass classification

K-NEAREST NEIGHBOURS (KNN)

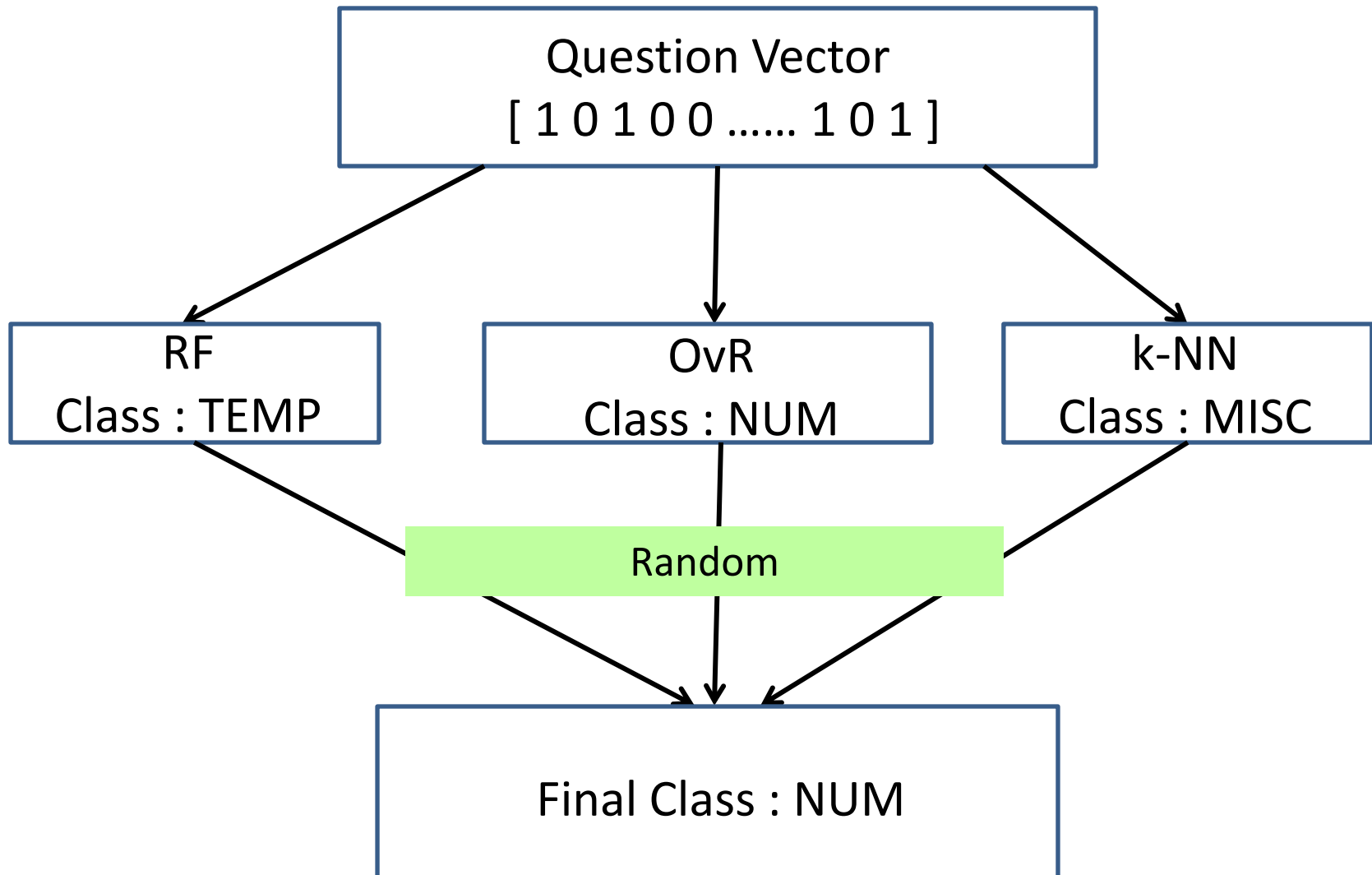


- Majority class vote of its neighbours
- Being a non-parametric method, it is often successful in classification situations where the decision boundary is very irregular
- Simple classifier

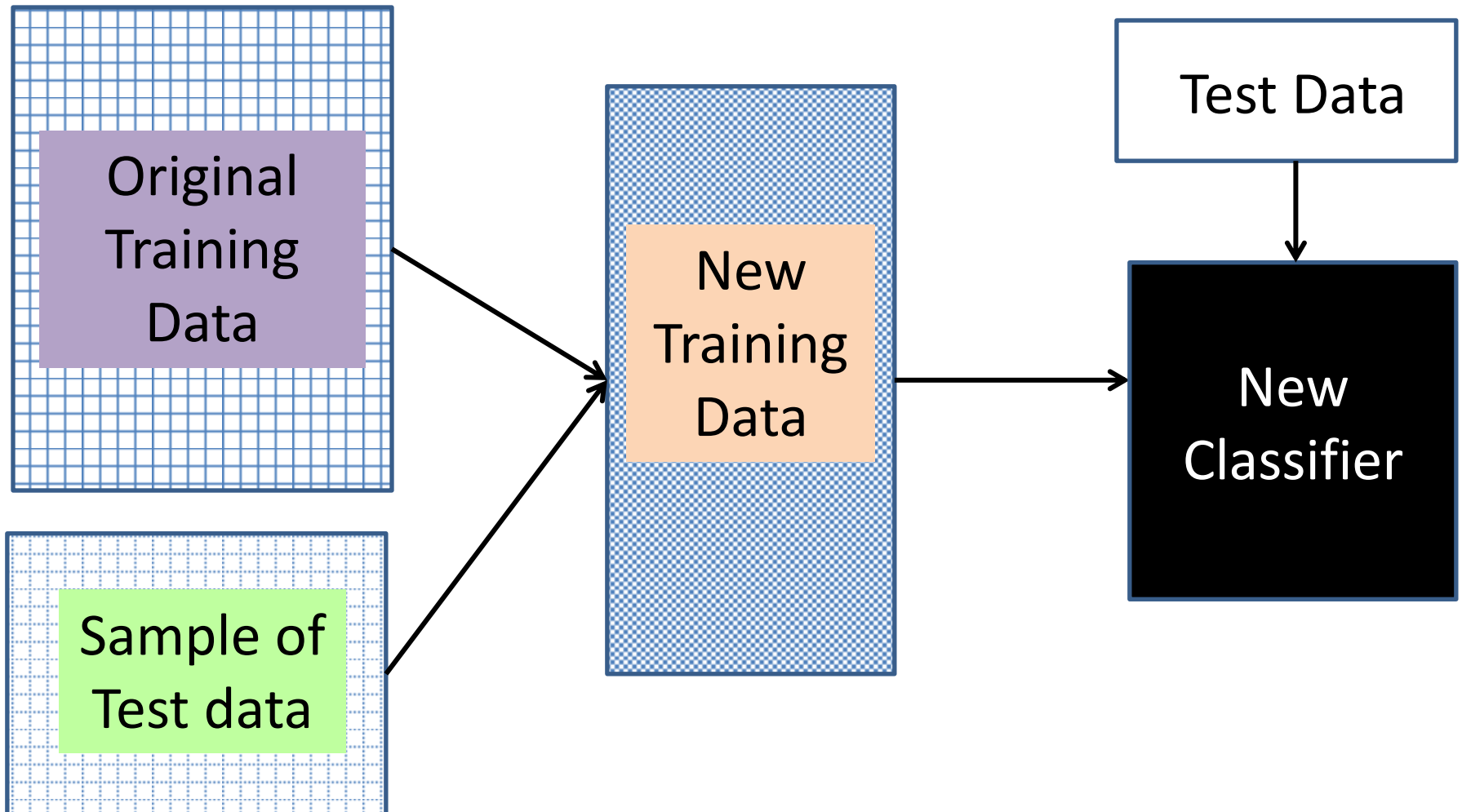
ENSEMBLE CLASSIFIER



ENSEMBLE CLASSIFIER



RETRAINING



DATASET

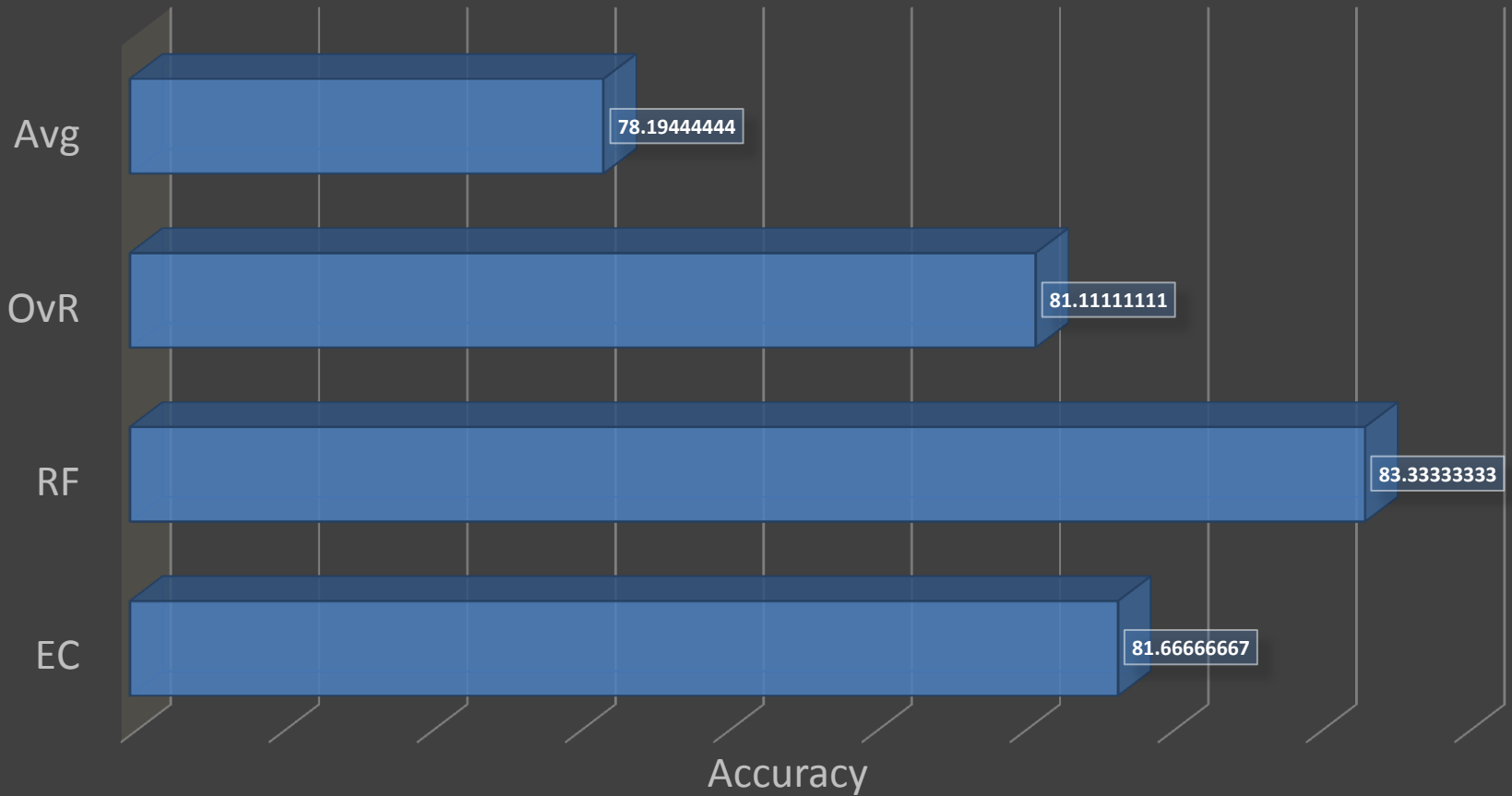
CLASS	NO. OF QUESTIONS
Person (PER)	55
Location (LOC)	26
Organization (ORG)	67
Temporal (TEMP)	61
Numerical (NUM)	45
Distance (DIST)	24
Money (MNY)	26
Object (OBJ)	21
Miscellaneous (MISC)	5

EXPERIMENTS

- *scikit-learn* toolkit of Python 3
- Training-Validation Split = 9:1
- No. of trees in RF = 100
- Classifier for OvR = Linear SVC
- No. of neighbours in kNN = 30

RESULTS

OVERALL PERFORMANCE



RESULTS

	I	IC	P	R	F-1	
PER	24	20	0.833333	0.740741	0.784314	EC
	25	21	0.84	0.777778	0.807692	RF
	23	19	0.826087	0.703704	0.76	OvR
LOC	26	21	0.807692	0.913043	0.857143	EC
	26	22	0.846154	0.956522	0.897959	RF
	26	21	0.807692	0.913043	0.857143	OvR
ORG	36	19	0.527778	0.791667	0.633333	EC
	34	19	0.558824	0.791667	0.655172	RF
	40	19	0.475	0.791667	0.59375	OvR
NUM	30	26	0.866667	1	0.928571	EC
	29	26	0.896552	1	0.945455	RF
	29	26	0.896552	1	0.945455	OvR
TEMP	25	25	1	1	1	EC
	25	25	1	1	1	RF
	25	25	1	1	1	OvR
MONEY	16	13	0.8125	0.8125	0.8125	EC
	16	13	0.8125	0.8125	0.8125	RF
	12	12	1	0.75	0.857143	OvR
DIST	20	20	1	0.952381	0.97561	EC
	20	20	1	0.952381	0.97561	RF
	22	21	0.954545	1	0.976744	OvR
OBJ	3	3	1	0.3	0.461538	EC
	5	4	0.8	0.4	0.533333	RF
	3	3	1	0.3	0.461538	OvR
MSC	0	0	NA	NA	NA	EC
	0	0	NA	NA	NA	RF
	0	0	NA	NA	NA	OvR

CONCLUSION & FUTURE WORK

- Machine learning algorithms for code-mixed Bengali-English data
- Scalable to other code-mixed questions since it is not language dependent
- Incorporate feature engineering – syntactic and semantic features
- Apply other ML algorithms
- Experiment with multi-script data

ACKNOWLEDGEMENT

This work is supported by the project

“To Develop a Scientific Rationale of IELTS (Indo-European Language Systems)

Applying A) Computational Linguistics &
B) Cognitive Geo-Spatial Mapping Approaches”

funded by the Ministry of Human Resource
Development (MHRD), India

ధన్యవాదాలు ధన్యవాద ఆమోదితనానికి కనాబాద
ధన్యవాదగళు మణి పర్యరూఢ యనవార ధన్యవాదాలు ఆమోదితనానికి కనాబాద
యనవార ఆమోదితనానికి ధన్యవాదగళు మణి ధన్యవాదాలు నునానికి
ధన్యవాద ధన్యవాదాలు నునానికి ఆమోదితనానికి కనాబాద పర్యరూఢ
మణి నునానికి **THANK YOU** ధన్యవాదగళు
యనవార شکر ధన్యవాదగళు మణి పర్యరూఢ యనవార కనాబాద ధన్యవా
ధన్యవాదాలు ధన్యవాద ఆమోదితనానికి కనాబాద ధన్యవాదగళు మణి
ధన్యవాదగళు మణి పర్యరూఢ యనవార ధన్యవాదాలు ఆమోదితనానికి కనాబాద

Any
Questions!