Communications Regulatory Authority State of Qatar

هيئة تنظيم الاتصالات دولة قطر

November 4, 2019 ICANN Tech Day

Machine learning for web content Classification

.QA use case for registered domains web content classification

By Mohaseenkhan Chinwal

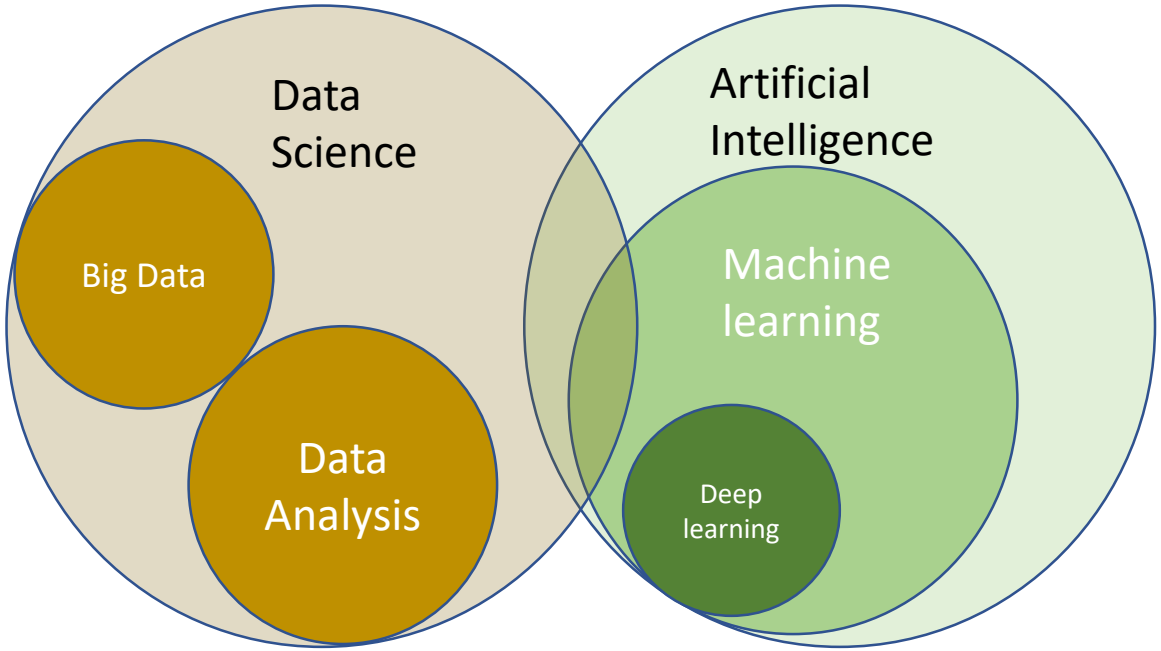
Table of contents

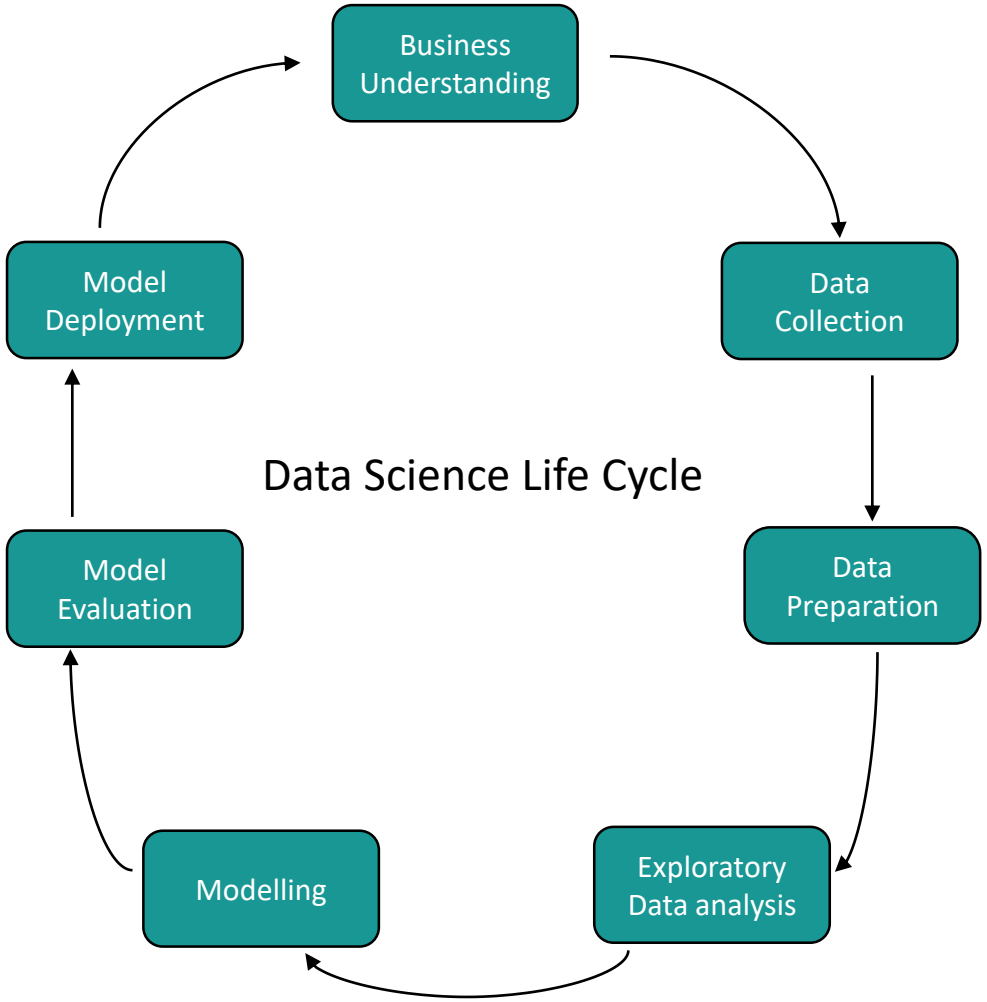
- Brief overview.
- Use case details as per Data science Life Cycle.
  - ✓ Business Understanding.
  - ✓ Data Collection.
  - ✓ Data preparation.
  - ✓ Exploratory Data analysis.
  - ✓ Modeling.
  - ✓ Model Evaluation.
  - ✓ Model Deployment.
- Scope of improvement.
- Use case work flow.
- Q/A.

The Big picture

- Problem Statement : Classification of web content of .QA registry domains in terms of business and non-business content.
- Solution: Building a mechanism to predict the probability of a website being hosted for either commercial or non commercial purpose.
- Expected Outcome: Study of domain usage patterns by the domain owners, analyzing the gaps between Registry-Registrar model and consumers in terms of domain registration and its use.
- Usage: The problem statement can be expanded (sub-scaled) on the classification front, which can be derived from the web content.

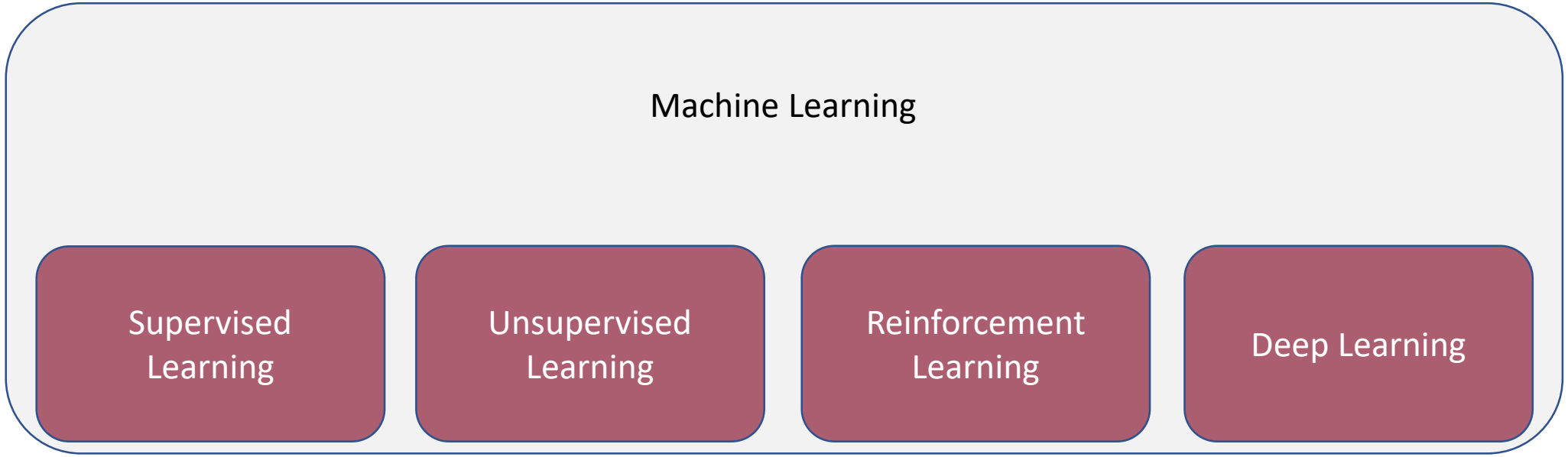
Overview of underlying technologies





Machine learning is branch of AI which deals with teaching machines to learn from experiences based on training them in terms of data. There are various types of machine learning techniques such as supervised, unsupervised reinforcement and deep learning.

- Supervised learning.
- unsupervised learning.
- Reinforcement.
- Deep Learning



Business Understanding:

DNS registry actively registered domain web-content sites segregation in terms of defined category by the business.

- Domain owners are encouraged/motivated/requested to use .com.qa for business use.
- Websites hosted for commercial purposes under .qa needed to be approximated for further study.
- This study helps us to enhance our process to reverse this trend.

Data Collection:

- Data was collected from various sources like data world, Kaggle and Data network, which has websites url lists segregated in various categories.
- A pre-classified data from these sources was used

Category (Labelled dependent value)	Text Content(Independent value)
Class_1(Commercial) (1)	Business and Industry. Finance. Shopping.
Class_2(Non-Commercial) (0)	Home and garden. Law and government. People and society. Pets.

Data Preparation

- Web scraping to collect text content of home page of websites.
- Removal of special characters, stop words, tokenizing words and letters case normalization.
- Stemming words using Lancaster algorithm
- Creating training dataset manually with keywords based segregation.

ID	URL	Strings
1	https://www.6sqft.com	What you need to know about Columbus Day an...
2	https://www.aerofarms.com	The Future of Farming? No Sun, No Soil, But...
3	https://www.durrell.org	Membership Cafes Latest Rewilding Islands M...

ID	URL	Strings	Class_1	Class_2	Word match count	Category
1	https://www.6sqft.com	real estate trends Greenpoint…..	80.0	20.0	10	Class_1
2	https://www.aerofarms.com	The Future of Farming?....	100.0	0.0	12	Class_1
3	https://www.durrell.org	Membership Cafes Latest Rewilding….	13.8	86.111	36	Class_2

Exploratory data analysis

Less Keywords Run1

	Class_1	Class_2	Total_matches
count	4540.0	4540.0	4540.0
mean	26.43	12.97	11.63
std	37.62	24.51	26.71
min	0.0	0.0	0.0
25%	0.0	0.0	0.0
50%	0.0	0.0	0.0
75%	60.0	16.6	11.0

Classes	Count
Class_1	1349
Class_2	438

More Keywords Run2

	Class_1	Class_2	Total_matches
count	4540.0	4540.0	4540.0
mean	23.29	17.59	13.97
std	34.34	29.03	33.41
min	0.0	0.0	0.0
25%	0.0	0.0	0.0
50%	0.0	0.0	0.0
75%	50.0	30.0	13.0

Classes	Count
Class_1	1149
Class_2	705

Most Keywords Run3

	Class_1	Class_2	Total_matches
count	4576.0	4576.0	4576.0
mean	22.5	19.09	16.07
std	33.4	30.33	40.40
min	0.0	0.0	0.0
25%	0.0	0.0	0.0
50%	0.0	0.0	0.0
75%	46.26	35.08	15.0

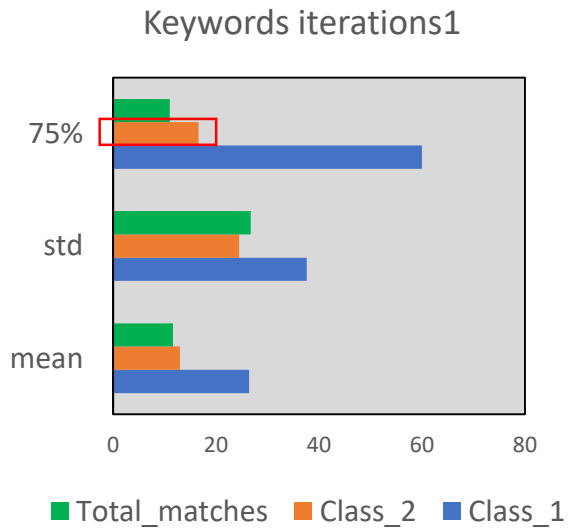
Classes	Count
Class_1	1089
Class_2	782

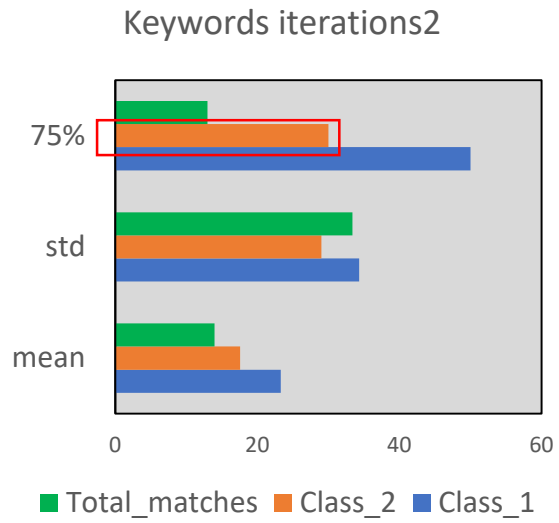
Exploratory data analysis

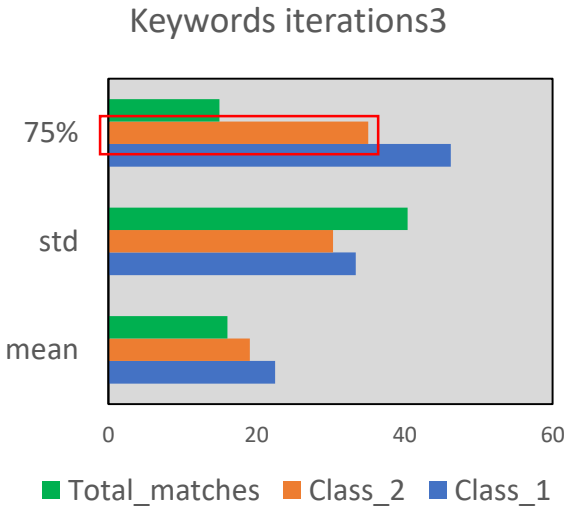
	Class_1	Class_2	Total_matches
mean	26.43	12.97	11.63
std	37.62	24.51	26.71
75%	60	16.6	11

	Class_1	Class_2	Total_matches
mean	23.29	17.59	13.97
std	34.34	29.03	33.41
75%	50	30	13

	Class_1	Class_2	Total_matches
mean	22.5	19.09	16.07
std	33.4	30.33	40.4
75%	46.26	35.08	15



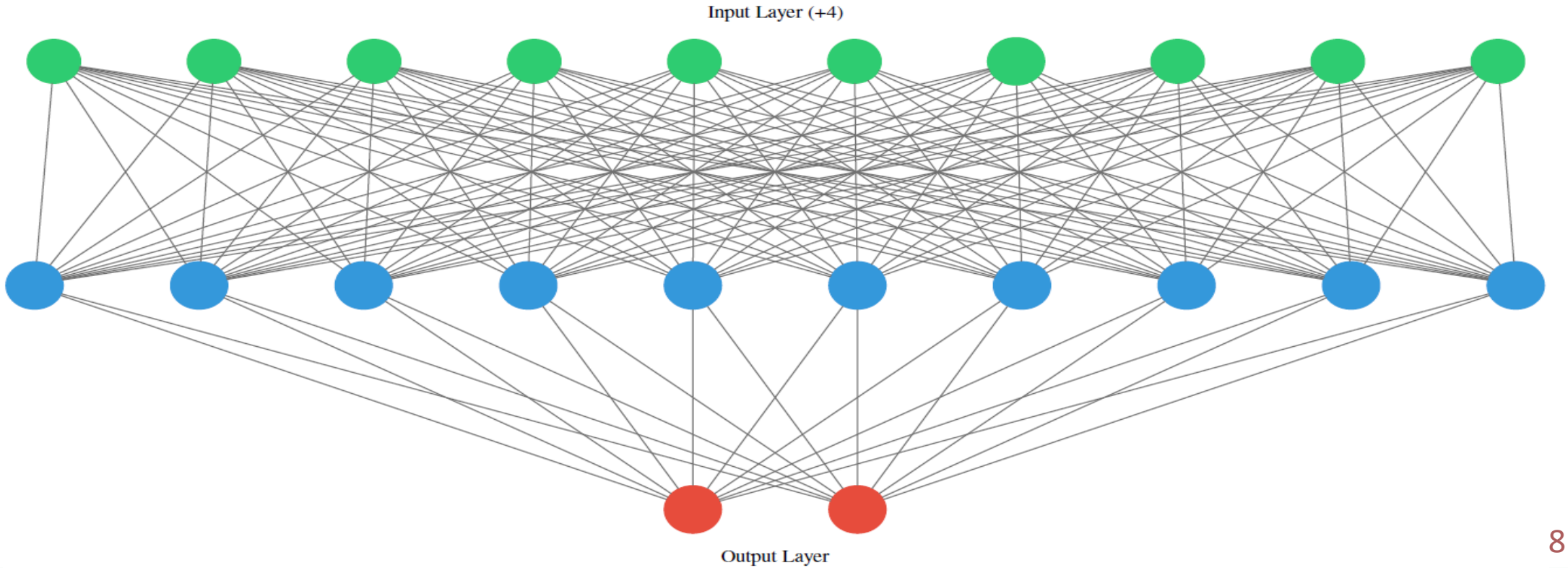




Modelling

- The designed models uses bag of words approach to feed data to neural network computations.
- Has one hot vector encoding. The inputs layer size equal to the word count of the entire corpus.
- It uses five models for aggregation with 10,000 iterations per model.
- One hidden layer with 10 neurons. Uses sigma function for squashing and its derivative for correcting weights during back propagation.

ANN Sample Visual for Qatar Domains Registry com.qa training dataset using Tensorflow, Keras and graphviz



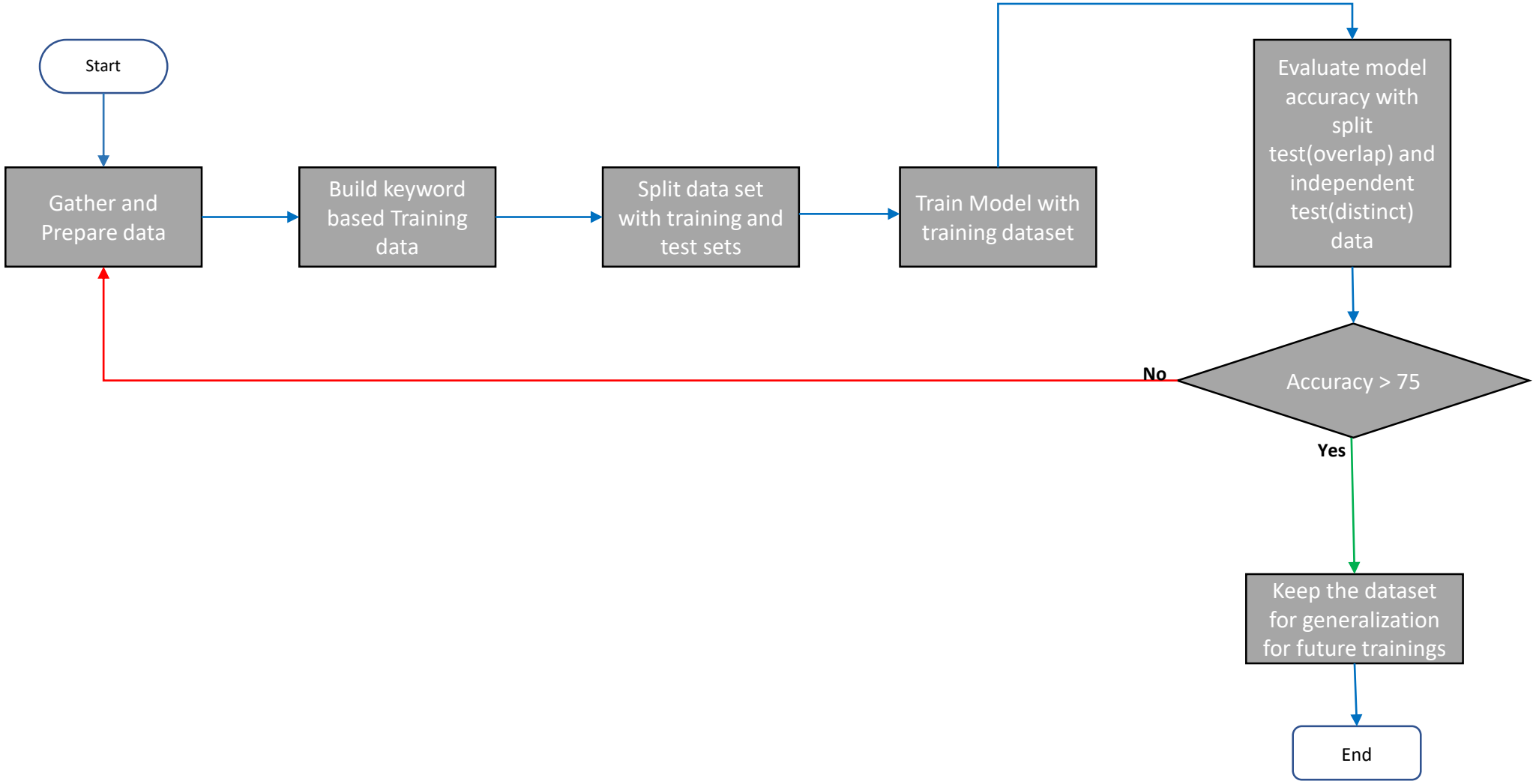
Model Evaluation

Test No	Model	Training document count	Vocabulary word count	Confusion Matrix	Model accuracy	Test data class samples	Training data split	Test-train-set-overlap
1	Model 1	50	5238	[[419 227] [1227 1472]]	0.5653	Class 2:646 Class 1:2699	25:class1 25:class2	overlap
2	Model 2	100	5759	[[370 276] [724 1975]]	0.7010	Class2:646 Class1:2699	50:class1 50:class2	overlap
3	Model 2	100	5759	[[27 23] [17 102]]	0.7633	Class2:50 Class1:119	50:class1 50:class2	distinct
4	Model 3	15678	61933	[[0 50] [0 119]]	0.7041	Class2:50 Class1:119	No split	distinct
5	Model 4	1450	51894	[[23 13] [82 58]]	0.46	Class2:36 Class1:140	725:class1 725:class2	distinct

Scope of improvement

- Use of reliable and accurate training data generation methods.
- Use of large training data in effective way to generalize the results of the model.
- Widening the scope of categories from dual to multiple labels.
- Use of better data encoding algorithms.
- Redesigning and Tuning ANN parameters.
- Use of better data parsing and cleansing techniques.

Use case work flow:



- *Web data extraction: (Stage 1) (prepare training data set)*

✓ The first part reads from list of domains, downloads the html content from its URL and extracts the text content of each domain which only contain English words.
✓ Outcome of this process to get the URL (domain name) links and associated web site home page text based keywords.

Output:

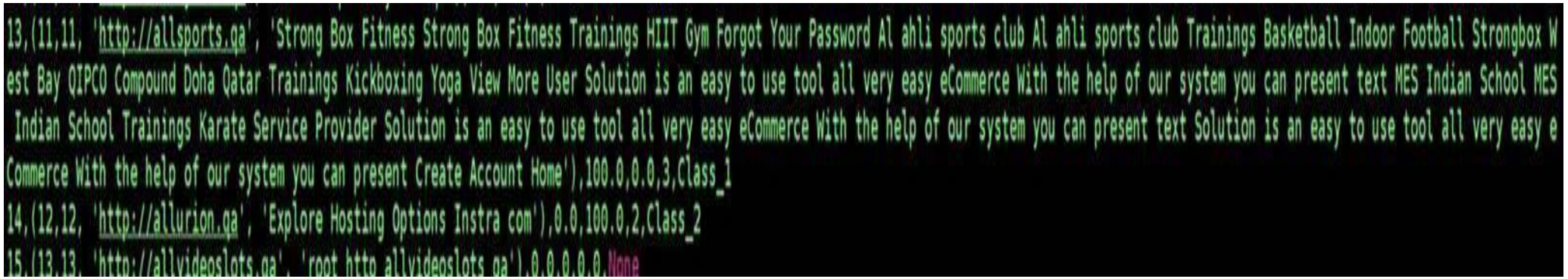
```
(0, 0, 'http://00.qa', 'International Trademark Registration International Domain Registration New gTLDs')
(1, 1, 'http://01.qa', 'International Trademark Registration International Domain Registration New gTLDs')
(2, 2, 'http://01telecom.qa', 'Partners Read More 974 4033 9170 info 01telecom qa Home Read More Careers Services Contact Us About Us')
```

- *Data classification: (Stage 2) (prepare training data set)*

✓ The second part takes the input from the previous module compares them with the predefined words and calculates the percentage of words falling into different data class. Outcome of this process is to get the url(domain name) and number of matching keywords in each class its percentages.

✓ The domains are allocated to a particular category (class_1,class_2 and none).

Output:



- *Producing training data set, running the training data on artificial neural network algorithm and writing the predicted numerical values on a file (synpase). (Stage 3) (train the model with training data set)*

✓ The is third part, uses the category , domains and text contents. The website text data relating to all domains in the list is broken-down into a word list.

✓ The word list is converted into numerical notations for ANN(Artificial Neural Network) to work.

✓ The next parts follows the mathematics needed to build the model, train it and write its predictions. Such as sigmoid, its derivative , data cleansing, think and activation functions.

✓ Training of the model with training dataset happens in this part.

✓ ANN algorithm starts learning with data given to which was manually produced in data classification and starts writing its learning to a synapse file. (Acts its memory to apply its learning to production dataset in our case .qa domain list websites)

Output:

- *With the production dataset(actual data whom we need to classify) we use stage1 and stage2 to generate the classification data(data in a format used by the ANN algo).*
- *Then we use stage4 to make predictions using the model on production dataset which is already trained on large data set in stage 3. (The accuracy of model completely depends on abundance of training data and text extraction from the web content)*

Output:

```
Confusion Matrix :
[[ 27  23]
 [ 17 102]]
Accuracy Score : 0.7633136094674556
Report :
              precision    recall  f1-score   support

           0       0.61      0.54      0.57        50
           1       0.82      0.86      0.84       119

    accuracy                           0.76       169
   macro avg       0.71      0.70      0.71       169
weighted avg       0.76      0.76      0.76       169
```

Q/A

THANK YOU