**Chapter 5**

**Result and Analysis**

**5.1 Analysis of Rule-Based Approach**

To evaluate rule-based approach we have done some analysis on the *Primary Corpus*. The findings of the analysis are discussed in the following sections. As the corpus is mixed with different types of sentences with high variations in spelling, words, subject matters, it was very difficult to analyse them under concrete set of rules.

**5.1.1. Analysis of Interrogative Corpus**

This corpus is the most important corpus for our experiments in rule-based approach. To analysis the sentences of the corpus we have calculated the average number of words, average length of words and average number of letters per sentences. The calculation is reported in TABLE II. The position of the Bangla question words serves a great importance in our rule-based approach. Therefore, we have made a Bangla question word list which contains 20 words. The list is depicted in Figure 2(Bangla Question Word List). We have calculated the position of these question words in the interrogative corpus. This calculation is reported in TABLE III. We have also seen that the word “*Naki*” is used as the question word in 24 sentences

Ki, Keno, Kivabe, Kothay, Koto, Kar, Kon, Kobe, Kisher, Kokhon, Ke, Kemon, Koy, Ke Ke, Kake, Kara, Kader, Koi, Koyta, Kotha.

1. Bangla Question Word List (figure 2)
2. Interrogative Corpus Analysis

|  |  |
| --- | --- |
| Average Number of Words | 4.7 |
| Average number of Letters per Word | 4 |
| Average Number of Letters | 19 |

a. All the averages denotes the average per sentence

1. Bangla Question word Position Analysis in interrogative corpus

| Position of Bangla Question Word in Sentences | Corpus Information | |
| --- | --- | --- |
| Number of Sentences |
| 1st Word | 112 | |
| Last word | 171 | |
| 2nd Word | 167 | |
| 3rd Word | 37 | |
| 4th Word | 4 | |
| 5th Word | 4 | |
| 6th Word | 2 | |
| 7th Word | 1 | |
| 8th Word | 1 | |
| Just Before the Last Word | 52 | |

## **Analysis of Other Mega Corpus**

In this corpus all the sentences without question mark remain present. Like the interrogative corpus, we have calculated the average number of words, average length of words and average number of letters per sentences for this corpus. The calculation is reported in TABLE IV.

1. other mega Corpus Analysis

|  |  |
| --- | --- |
| Average Number of Words | 5 |
| Average number of Letters per Word | 4 |
| Average Number of Letters | 11 |

* 1. All the averages denotes the average per sentence
  2. **Making of Dictionary**

We have made the list of words that appeared in the corpora. The dictionary made from the words have depicted the variations found in the spelling of the transliterated words as there is no standard form or rules. This varies from people to people.

Example:

The word “আপনি” is found as follows:

-apni

-aapni

-apny

-apne

## **5.3 Evaluation of Rule Based Approaches**

From the set of rules prescribed in section 3.1, we have tested the interrogative corpus. The main basis of this approach is the position of the Bangla question words in the sentence. From TABLE III, we observe the positions of the Bangla question words. Excluding the last position for the other position we find the mean position of the Bangla question words which is 1.90.

At first, we have tested the corpus according to the rule 1, that is the presence of Bangla question word at the 1st or last position of the sentence then gradually we integrated the other rules and observe the combined effects of the other rules in the improvement of accuracy. The evaluation is given in TABLE V.

1. Rule based Approach Evaluation

| Method | Accuracy % |
| --- | --- |
| Rule 1 | 40.42 |
| Rule 1 + Rule 2 | 64.29 |
| Rule 1 + Rule 2 + Rule 3 | 71.71 |
| Rule 1 + Rule 2 + Rule 3 + Rule 4 | 75.14 |

From this level of accuracies from a known set we can say the rule-based approach is not satisfactory. As the number of rules can’t be made concrete due to the diversity and varieties of sentences. From TABLE (position of wh word), we have seen that the maximum likely position of Bangla question word is the last word of the sentence. In the sentence “*Tader jete bollam kothay ar tara gelo kothay*”, the question word “*Kothay*” is at the last position but the sentence is not an interrogative sentence whatsoever.

From this, we say that the rule-based approach is not pragmatic and there should be a learning based approach for the identification.

* 1. **Analysis of Learning-Based Approach**

We have analyzed the secondary corpus and the tertiary corpus with the help of SVM, k-NN, MLP and logistic regression classifiers. The insights we have found for the classifiers show us the significance of the problem and our approach of solution. At first, we have used the secondary corpus where the training and test set have come from the same domain. Then we have taken the tertiary corpus where the training and test set differs in terms of subject domain.

**5.3.1 Analysis of Secondary Corpus**

To evaluate the secondary corpus, we have taken 30% of the data as the test set and the rest as the training set. Though the domain is same, it has huge variations as it is taken from real people. Few sentences from this corpus is given in figure 3. We have labeled the test set and the training set then we have found out the accuracy. The training time was really small.

-keno jeche nijer desher nam kharap korchen apnara

-Bangladesh kar sata khalba

-Ajke brishti houar shomvabona ache ki?

- ami tomer sathe akmat

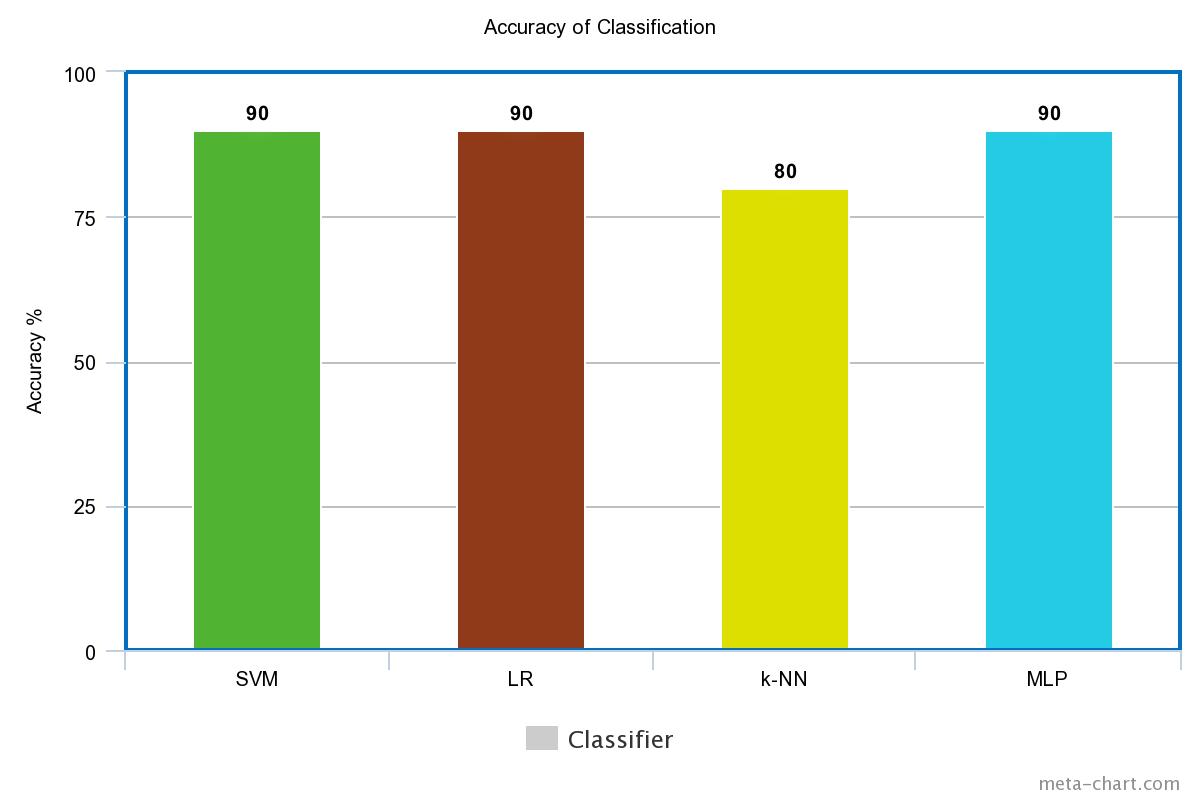
- khela kon chennel a dakhay ?

1. Sentences in Secondary Corpus (figure 3)

The result of the experiment was quite satisfactory and significant. The accuracy level we have observed is as follows:

* SVM : 90.3599755949 %
* Logistic Regression: 90.176937156 %
* K neighbors Classifier: 80.65893 %
* MLP classifier: 90.3599755 %

These findings are depicted in Figure 4.



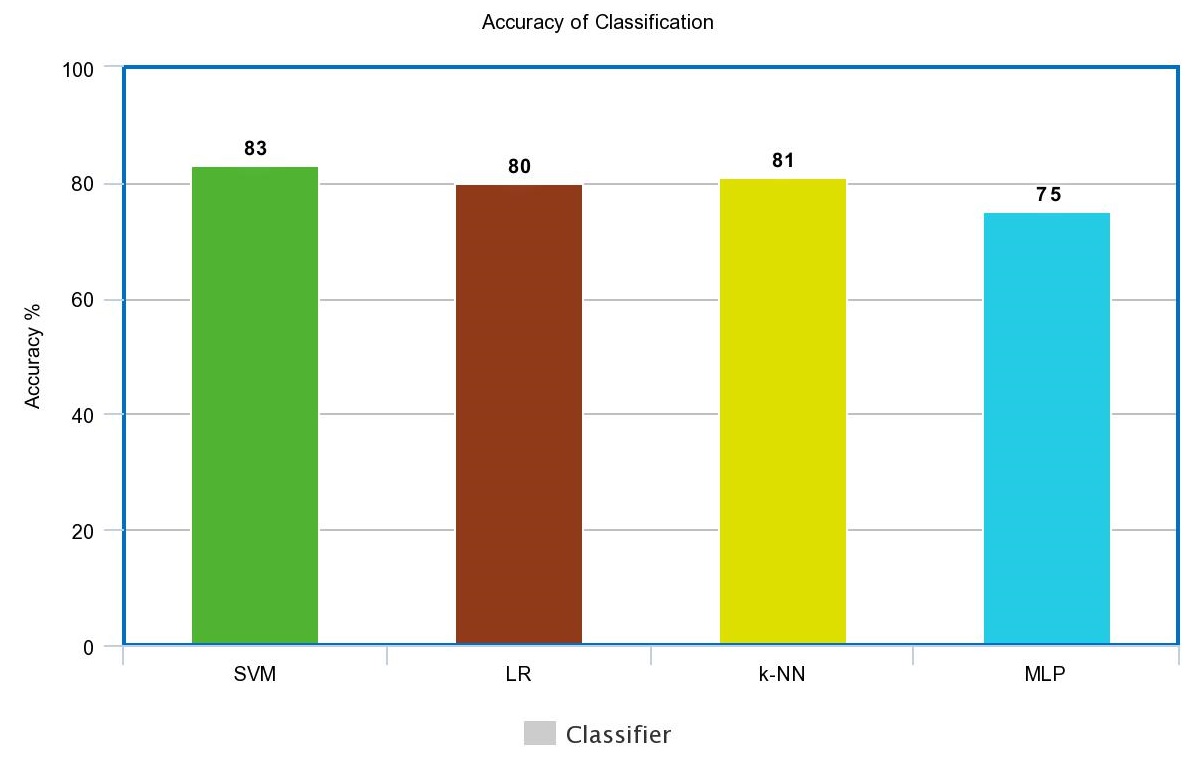
**5.3.1 Analysis of Tertiary Corpus**

The tertiary corpus contains data from two domains, one is cricket another is hand picked data for admission test related queries. We have implied the same four classifiers for this corpus. This time admission test based dataset was the training set and the cricket based dataset was the test set.

The result we have found are as follows:

* Accuracy from SVM: 82.6441927838
* logistic regression: 79.79984198051093
* KNN: 80.58993942586252
* MLP: 74.66420858572558

These findings are depicted in Figure 5.



**5.4 Evaluation of Rule-Based Approach:**

We observe that the accuracy when we have used the secondary corpus differs from the findings of the teritary corpus. It happened due to the change of the domain of the training dataset. As the admission dataset contains formal sentences and sentences with standard form of transliterated sentences, the accuracy label drops. The test dataset is pretty raw in nature. With a standard form of training set our model of classification will work swiftly but with the standard training dataset the real time scenario will be ignored. People use different styles, spelling, words for writing in transliterated form. Therefore the test dataset should address this nature for signifying the pragmatic approach of detecting interrogative transliterated bangla sentences. The drop in accuracy also happened as the number of sentences in training dataset has decreased. Keeping the domain of the training and test dataset will naturally show good results in classification as we are using CountVectorizer method for feature extraction.

The accuracy of k-NN remains almost same. 80% and 81% for secondary corpus and tertiary corpus respectively. This happens because k-NN does not depend on the labeling of the training dataset. It just finds out the “k” neighbors of the test datapoint. With the increase of the interrogative sentences in the tertiary corpus the accuracy of finding the nigbors for such sentences increases therefore for tertiary corpus only k-NN works better than other classifiers comparing to the secondary corpus evaluation.

* 1. **Analysis of the Deep Learning Approach**

To implement CNN, we have used our tertiary corpus as this corpus contains the most number of sentences.For training, we have set the parameters as follows:

Dimensional Embedding = 128

Number of filters per filter size = 128

Filter size = 3, 4 and 5

Dropuout = 0.5

We have taken 10% of the data as test dataset. After the training the summary can be visualised by graphs using TensorBoard.

The loss and the accuracy of the summary is depicted in Figure 5. Red represents the training data and blue represents the test data.



Figure 5: Accuracy of the CNN

The insights we have found from this experimet can be described by the following points:

* The model is overfitted as the test data accuracy is bellow than the training data accuracy. By increasing the dataset and making the regularization stronger we can overcome this situation.
* After a significant ammount of training steps (more than 30,000) we observe an accuracy level of 85.6% for the test data.
* If we use pretrained wor2vec vectors for the embedding matrices we can gain more accuracy from this approach.