NLP Assignment

By: Aryan Chaturvedi

B.Tech. Pre-Final Year

IIT Roorkee

[Aryan\_c@ch.iitr.ac.in](mailto:Aryan_c@ch.iitr.ac.in)

**Text Classification Model**

* Dataset Used:
  + **IITJEE NEET AIIMS Students Questions Data**

The dataset contains **Students-questions.csv** file in version 1 as of now.  
Inside the CSV file, we have two columns:

* eng: The full question or description of the questions
* Subject: Which subject does the question belong to. It has 4 classes, **Physics**, **Chemistry**, **Biology**, and **Mathematics**.

So, it's basically an NLP problem where we have the question description, and we need to find out which subject does this question belongs to.

Link To dataset: <https://www.kaggle.com/datasets/mrutyunjaybiswal/iitjee-neet-aims-students-questions-data>

A screenshot of a computer

Description automatically generatedThe dataset contains total 120000+ questions from which I have sampled 10000 questions.

In this task I have used three different approaches to find the best classification score:

* Approach 1: Using TF-IDF and Machine learning Classifiers.
* Approach 2: Using Pre-trained BERT model based embeddings of questions as features for different Machine learning Classifiers.
* Approach 3: Fine Tune BERT Based Classifier on this question dataset.

**Approach 1: Using TF-IDF and Machine learning Classifiers.**

**Code File:** <https://github.com/AryanChaturvedi/TextClassification/blob/main/Tf_Idf_Model.ipynb>

* Performed Basic EDA
* Cleaned texts from special characters using Regex
* Applied Basic NLTK Pre-processing like stemming, lemmatization and stop word removal.
* Created a Document Term matrix using TF-IDF
* Applied Different ML Classification Models to classify questions
* Result is summarised here:

|  |  |
| --- | --- |
| Model with TF-IDF | Accuracy (%) |
| KNN Classifier | 51 |
| Random Forest | 77 |
| Naive Bayes | 86 |
| Logistic Regression | 88 |
| SVM | 89 |

**Approach 2: Using Pre-trained BERT model based embeddings of questions as features for different Machine learning Classifiers.**

**Code File:**

<https://github.com/AryanChaturvedi/Text-Classification/blob/main/BERT_Classification.ipynb>

* Performed Basic EDA
* Cleaned texts from special characters using Regex
* With pre trained BERT (bert-base-uncased) model generated sentence embedding using average of 2nd to last layer embeddings.
* Applied Different ML Classification Models on these embeddings as a feature vetors
* Result is summarised here:

|  |  |
| --- | --- |
| Model with Bert Embedding | Accuracy (%) |
| KNN Classifier | 44 |
| Random Forest | 86 |
| Logistic Regression | 80 |
| SVM | 31 |
| XGBoost | 88 |

**Approach 3: Fine Tune BERT Based Classifier on this question dataset.**

**Code File:** <https://github.com/AryanChaturvedi/Text-Classification/blob/main/BERT_Classification.ipynb>

* Performed Basic EDA
* Fine Tuned BERT Sequence Classification Model
* Improved accuracy to **91.35%** after fine tuning.

Overall best results are summarised as below:

|  |  |
| --- | --- |
| Model | Accuracy (%) |
| KNN Classifier with TF-IDF | 51 |
| KNN Classifier with BERT Embeddings | 44 |
| Random Forest with TF-IDF | 77 |
| Random Forest with BERT Embeddings | 86 |
| Logistic Regression with TF-IDF | 88 |
| Logistic Regression with BERT Embeddings | 80 |
| SVM with TF-IDF | 31 |
| SVM with BERT Embeddings | 89 |
| Naive Bayes with TF-IDF | 86 |
| XGBoost with BERT Embeddings | 88 |
| Fine Tune BERT Classifier | 91.35 |

Detailed Procedure of each method is mentioned in notebook.

My Resume: <https://drive.google.com/drive/u/0/folders/10bCFdCaYAysUiMbjNa2wXyegwOpg4ftv>

My GitHub Repo: <https://github.com/AryanChaturvedi>

My Medium Blogs on NLP: <https://medium.com/@aryan_c>

My Blogs:

* [*Sentence Pair Classification using BERT embedding*](https://medium.com/@aryan_c/sentence-embedding-with-sentence-transformers-sbert-part-1-1542ba42a65d)
* [*Fine-Tune BERT for Domain-Specific Embeddings*](https://medium.com/@aryan_c/sentence-embedding-with-sentence-transformers-sbert-part-2-99439520462e)
* [*Sentence Transformers: Beginners Guide*](https://medium.com/@aryan_c/sentence-embedding-with-sentence-transformers-sbert-part-3-90365023b74f)
* [*SBERT and SimCSE: Practical Implementation with Cookbook*](https://medium.com/@aryan_c/sentence-embedding-with-sentence-transformers-sbert-part-4-162969d4ec9e)