**Advance Natural Language Processing**

**Sematic text classification of Covid-19 Data.**

Team-mates

1. Jyothi Pudota
2. Pravarsha Vantipalli
3. Likhita Perala

Overview: Our goal is to Investigate COVID-19 Q&A datasets for

chatbot training using Topic modeling. We provided a framework to preprocess users’ queries and retrieve either a conversational platform or paper references related to COVID-19. To do such a task, we introduce the following architecture:

* The user passes their question query, and the inputs are passed on to a classifier. We label 0 as general chat bot questions and label 1 as covid questions.
* The intent classifier would classify data based on the class of the given questions to pass into a knowledge model (class 1) or small talk (class 0). Depending on the class being labeled, the architecture retrieves the answer by passing into our knowledge-based model:
  + For Simple talk questions, it categorizes the question with the respective intent to retrieve one among all answers.
  + For the knowledge-based model, it retrieves related papers that fit with the topic domain using topic modeling.
* Based on the answer, the user will response and the architecture takes that query response back to the intent classifier again. The process follows again from the steps.

Data processing: We collected over 839 data points that split evenly for both covid-19 conversation (class 0) and paper-related questions (class 1). We preprocessed the questions from the available 2019 faq-covidbert dataset and normal conversation dataset. As words data are not passed directly but are passed into embeddings before going into the model, we pre-processed the dataset by using the following embedding techniques: Count Vectorizer and Td-Idf method.

Model setup: Here we propose three NLP (Natural Language Processing) models for the intent classifier: SGD classifier, Naïve Bayes, and Logistic classification. All models have the same pre-processing steps, and we focus only on the classification ones.

Pre-trained model: We split the data into a training set and a testing set with 4:1 ratio as 629 for training and 210 for testing. After preprocessing data, we pass the training data into all three models.

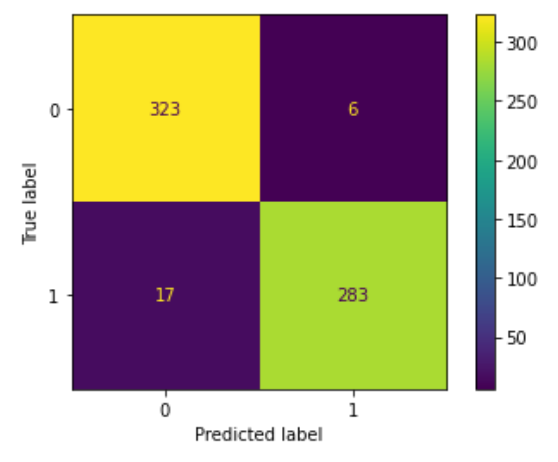
Performance metrics: There are four of them: Leave-one-out cross-validation (LOO CV), F-1 score, AUC score, and confusion matrices. LOO is evaluated on the entire dataset, F-1 score, and AUC score are evaluated on the testing set, and confusion matrix is evaluated on both the training and testing sets.

For example, we group the COVID-19 questions related to children into one intent and related to K-12 education into another intent. To provide a rich answers database, all our 839 data questions are classified into multiple groups of intents and the corresponding answers are also provided.

Knowledge-Based model: The Knowledge-based model is a topic modeling framework that searches for related papers with the user’s request. When a question is passed into the model through an UI interface, the model learns the topic embedded in that question to pass a suitable paper on the UI interface. The model is integrated with the web development team for scalability.

**Result**: Here we provide the results of three models based on all four mentioned metrics. SGD and Logistic Regression have high LOO as 0.87 and 0.86, so they are suitable to integrate into the intent classifier. The AUC score is 0.86 and near to 1, suggesting that they avoid overfitting up to 86%.

**SGD-Classifier**

Chart, line chart

Description automatically generated

Naive Bayes Classifier:

Chart, treemap chart

Description automatically generatedChart, line chart

Description automatically generated

Logistic Regression:

Chart, treemap chart

Description automatically generatedChart, line chart

Description automatically generated