Project Report

For the following project I will use the cross industry standard process for data mining (CRISP DM) it is the industry standard process for executing any data science project .

It has four step as listed bellow and I will add some other steps to handle this project since it is an NLP task.

Those are the step I will be using in the following project .

Problem Understanding

In this step we will be explaining the business problem statement as a data science problem and verify if we have a clear understanding of the problem.

Business Description (Company Description)

Translator Without Border is a Non profit organization which aims to build a word where there is no language barriers.

Recently they built a multilingual chatbot using cutting edge NLP techniques that aims to provides up-to-date information and responds to the concerns of marginalized language speakers as they ask questions in an open format.

Problem Statement TWB would like to incorporate an open dataset on Covid chatbot strings into their own chatbot models but the data doesn't match 100%.

How would you approach this challenge and align the topics with the various data points, or minimize the work of the content team who would need to do it manually. TWB's own chatbot training data is listed in the attached excel.

Link to the open covid chatbot strings dataset

The topic structure in TWB's internal chatbot data needs to be kept intact since they are mapped with their answers already. The aim is to expand the training set of each topic with the strings in the new open dataset where possible.

Also, new topics could be created from those data points that do not match with any.

Your task involves simplifying this process so that TWB's linguists do not have to manually classify each string one by one.

My Problem Understanding TWb have a dataset of intents and they corresponding sentences and would like to add more data to it.

They got a new dataset from Transperfect however the intent from that new dataset does not match 100% with the Twb dataset .

The problem is how to map sentences from the Transperfect dataset to Twb dataset to update it and if there are new intent that comes from that dataset how can we update the new dataset from TWB with those new intent that comes from Transperfect dataset .

Solution

First Approach to handle this problem.

From the first view this looks like an hybrid problem of both text classification and clustering problem.

Classification Problem we can consider our twb dataset as the train dataset because we can use to train a classifier that will take a sentence as the input and predict his intent as the output .

The dataset is already labeled , with the mapping of each sentence to the corresponding intent.

This is is a multi-class classification problem where we have n + 1 class where n is equal to the number of intents we have in the twb dataset and to that we add one more intent which is the unknown intent. It will contains all the sentences that are not in any other dataset.

Clustering problem For all the sentence which the intent is unknown we can run a clustering algorithm or topic modeling algorithm to discover what are the different topics or intents we can found in them and from that we can construct new intents.

Technical approach to handle the problem

As said in the introduction of this report I will be using cross industry standard process for data mining

Data understanding The next step in this problem is to understand our dataset and what is it about and if can find any tips to solve the problem from the dataset itself .

In this step we will check how big is our train dataset and our test dataset this will allow us to find the best machine learning model to use for the problem.

Data Cleaning We know from the experience that this is the most important phase in a data-science project since the quality of the output depend on the quality of the input.

For this NLP project here are the steps we will undergo in this part.

- The first step will be tokenization : this will consist of splitting each sentence into token, or word

- Stemming or lematization will consist of replacing each word or token with their root word or word they derived from.

Eg : word such as catching , caught will be converted to catch

- We can also remove stop word : word like `the` `this` or `a`should be removed to make the process smooth and have accurate results
- And also checking if the are word that are misspelled and fix them : After a first look into the Transperfect dataset we can see that word such as

They writes `corona` as `coronavirus` and sometimes as `corona virus`

The also writes `covid` as `covid` or `covid-19` or `covid19`

Text representation: Since computer cannot understand text we will convert our text to a format a computer can really understand and that format is number or vector.

For to perform this task we will generate our vocabulary which we have all the word we have in our dataset.

Once we have our vocabulary we will encode our word as one hot vector in that vocabulary.

We can use also an embedding layer to get an accurate representation of our word.

Once we are done with our text representation we can now perform the machine learning metrics.

Machine learning

- multi-class classification As said before this is multi-class classification we will use machine learning algorithm to handle this type of problem.

Due to small sample of our train dataset we will not urge to use deep learning technic we will try simple models like :

- Naive Bayes
- Logistic Regression (Softmax Regression)
- Support Vector Machine for mutli-class classification
- Or decision tree

If those models does not perform well we can try use to use neural network

The first approach to this task will be to use a simple Recurrent Neural Network without attention mechanism since we are dealing with short sentences.

Or For fun we can try models with attention mechanism such a LSTM to see how they will perform on this task.

Since we are dealing with a small dataset I am suspecting those deep learning models to overfit this dataset.

- unsupervised learning

In this step we can use clustering technic to find new intents in the dataset :

- K means clustering can be usefull here ...
- Or LDA Latent Direclet Annodation for topic modeling

Evaluation

Once our models are trained we will evaluate them on a small sample of test and the evaluation metric we will be using is the LogLoss

We can also use a confusion matrix

And since our dataset is relatively small we can use cross validation technic to evaluate our model. Leave on out cross validation can be useful here , or k fold cross validation.

After the evaluation we can deploy our model and put it in production.

Deployment

Once were done with the training and we selected the best model we can put it in production so that translator and linguists can use it to find intent for the new question they got from any data source.

The model can be deployed as a simple flask app where it will be use by translators.

The front end should have a simple form where a user can put a sentence and then the front end call a web api which will return the correct intent for the sentence.

We can also add some rating form where translator will rate the intent prediction from the scale of one to five and we can use this rating to improve our model later.

We can also add the model into a chatbot pipeline where each time a new sentence comes from users we can predict it's intent directly using the model and return the response to him according to the detected intent.