Capstone Proposal

“Comment Adiviner”

Comment label predictor

Domain background

The project’s domain traces back to a Customer Experience company that wants to feed from customer’s feedback and to provide their clients with customer insight. They want to somehow predict the topic of a customer’s comment.
Based on the comment’s characteristics, a supervised approach would be more convenient as the comments aren’t long enough to take a unsupervised approach.
Now, comments are being manually labeled and sent to the client. They spend too much time labeling comments, therefore, clients doesn’t have their customer’s feedback until much later when it might be too late to improve the customer experience.

Problem statement

The company can not provide in real-time the issues that the comments are attend to. Comments must be manually reviewed everyday and labeled by a person, the amount of comments per day is too big and can’t be handled anymore by people. They need to provide real-time labeled comments so their clients can understand which are the main negative comments on specific topics right now.

Datasets and inputs

The dataset used are comments from different hospital’s clients in Spanish language from 2017 to 2019 obtained through a tablet terminal where the customer’s are asked how was their experience at the Hospital.
The Dataset contains the Name of the hospital, the comment and the topic of the comment. I will use this data to train a model after preprocessing some details.

Solution statement

Using a supervised machine learning algorithm and text analytics I would train a model with labeled comments to predict the label of each comment based on the the principles of TFIDF (short for term frequency–inverse document frequency). TFIDF measures not only the term count of a comment, but also the frequency compensating the most and less used terms along the data frame. This will create a matrix of terms and their respectively label.

Benchmark model

Vector models proven very efficient in these kind of excelsis such as LinearSVC and SGDClassifier.

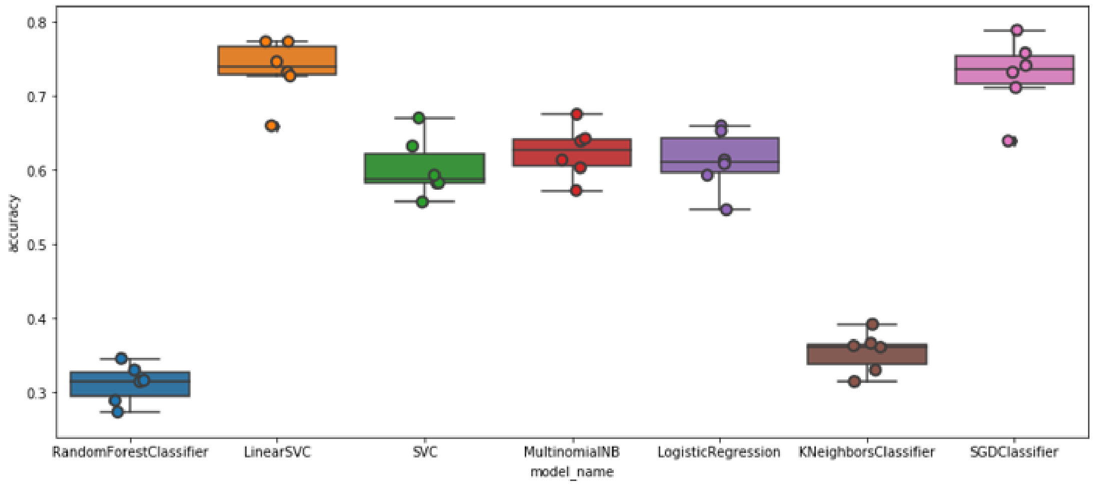
I believe a vector based algorithm is a good approach for this issue as it will transport a matrix of term frequency vectors to a 2 or 3 dimensional plane where the comments with similar term frequency of words would be close to each other and share the same label.

Evaluation metrics

The solution for the problem is to predict a real-time comment's topic/label related to hospital affairs such as (Food, Waiting Time, Stuff, Cleaning, Facility etc.).
This is important for reporting as they can update their client with categorized comments about what's wrong in the hospital and which comments are the most frequent. Therefore, the metric needed to evaluate the performance of the model must be the "Recall" and "Precision", as the classes are unbalanced "Accuracy" wouldn't work for our model. We need to make sure that the models predicts most of the comments correctly and with high precision or error in each Label.

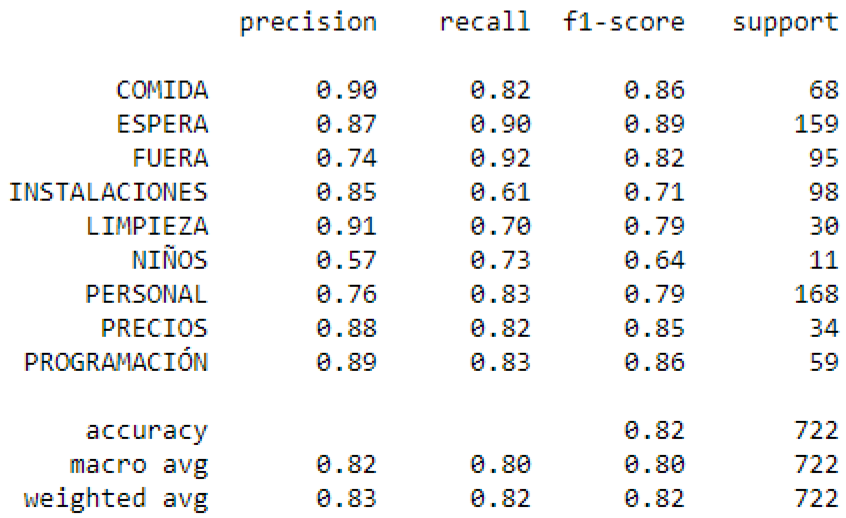
Project design

1. To processing the text (removing punctuation, stop words, lemmatization and stemming of words and vectorizing using TFIDF).
2. Compare different Sklearn algorithms and see their accuracy on our train and test data: SGDClassifier, LinearSVC, MultinomialNB, Logistic Regression…



3. Calculate the accuracy of the model on different n-grams distributions to create the training vocabulary. Depending on the length of comments a small n-gram could benefit the performance.
4. Train the selected model. LinearSVC.
5. Visualize test predictions in different formats to evaluate results.
6. Deploy model in API and create Lambda to receive input, which can be inputted in two different ways:
    a. by typing a single comment using an API with a lambda function or
    b. by uploading a CSV with comments using an API as well.

Finally, receive as output the comment's label or a csv with the predicted label and the comment itself.



```
              precision    recall  f1-score   support

      COMIDA       0.90      0.82      0.86        68
      ESPERA       0.87      0.90      0.89       159
       FUERA       0.74      0.92      0.82        95
INSTALACIONES       0.85      0.61      0.71        98
    LIMPIEZA       0.91      0.70      0.79        30
       NIÑOS       0.57      0.73      0.64        11
    PERSONAL       0.76      0.83      0.79       168
     PRECIOS       0.88      0.82      0.85        34
 PROGRAMACIÓN       0.89      0.83      0.86        59

    accuracy                           0.82       722
   macro avg       0.82      0.80      0.80       722
weighted avg       0.83      0.82      0.82       722
```