Machine Learning Lab Evaluation 2

Parameshwari S - CB.SC.I5DAS18026

April 04, 2021

Introduction

Code review is a technique of systematic examination of a code change. It is an important practice in software engineering. This is done to improve software quality by reducing defects. But this requires developers' knowledge about the code change and the context of the system. Understanding the code change and its context is one of the main issues reviewers face during a code review. Communicative intentions are one of the reasons for confusion related to the developer dimension of code reviews.

Problem Statement

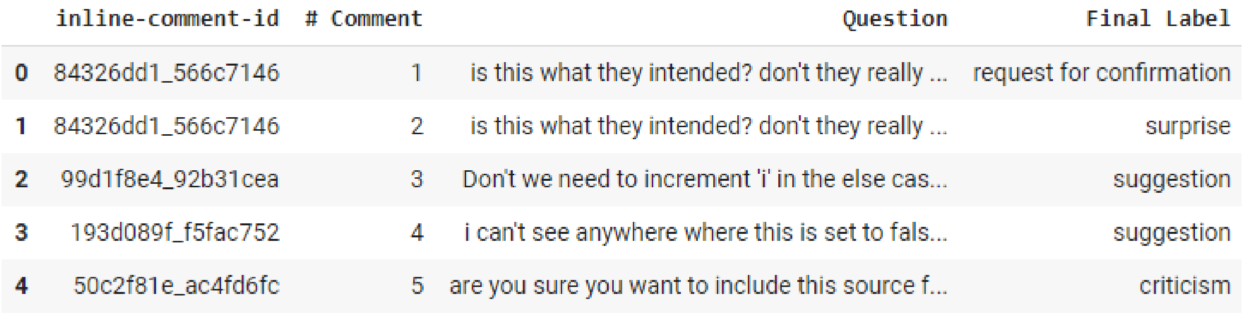
Task - Classification of code review questions.

Develop a classifier to predict the intention in the code review questions.

Dataset Description

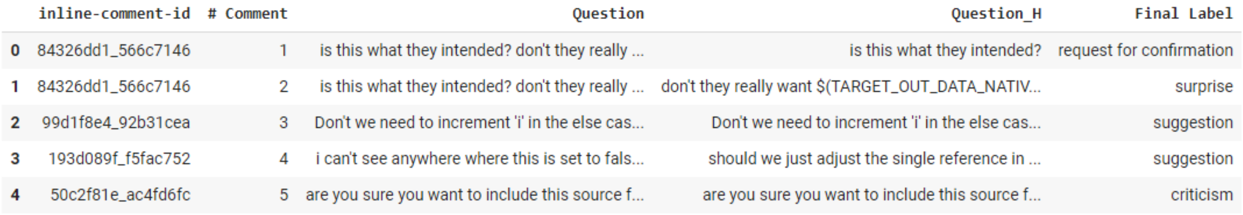
Dataset contains 499 rows and 4 columns:

- ***Inline-comment id -*** Id from which the comment was made
- ***Comment -*** Comment number
- ***Question -*** Comments made
- ***Final label -*** Type of comments (Suggestion, Requests, Surprise, Anger, criticism etc.)



	inline-comment-id	# Comment	Question	Final Label
0	84326dd1_566c7146	1	is this what they intended? don't they really ...	request for confirmation
1	84326dd1_566c7146	2	is this what they intended? don't they really ...	surprise
2	99d1f8e4_92b31cea	3	Don't we need to increment 'i' in the else cas...	suggestion
3	193d089f_f5fac752	4	i can't see anywhere where this is set to fals...	suggestion
4	50c2f81e_ac4fd6fc	5	are you sure you want to include this source f...	criticism

After analysing the dataset, I have manually edited it by making the highlighted text as a new column ***Question_H***. As many rows have the same text with different Final label, I found that Final label changes with respect to the highlighted text, hence I made them into a new column. This was done to check how accuracy changes. The new dataset looks like :



	inline-comment-id	# Comment	Question	Question_H	Final Label
0	84326dd1_566c7146	1	is this what they intended? don't they really ...	is this what they intended?	request for confirmation
1	84326dd1_566c7146	2	is this what they intended? don't they really ...	don't they really want $(TARGET_OUT_DATA_NATIV...	surprise
2	99d1f8e4_92b31cea	3	Don't we need to increment 'i' in the else cas...	Don't we need to increment 'i' in the else cas...	suggestion
3	193d089f_f5fac752	4	i can't see anywhere where this is set to fals...	should we just adjust the single reference in ...	suggestion
4	50c2f81e_ac4fd6fc	5	are you sure you want to include this source f...	are you sure you want to include this source f...	criticism

Feature Extraction

For this application, few columns are not required as they don't contribute much, hence have less effect on this application. ID of the person ***(inline-comment-id)*** and comment number ***(# comment)*** are irrelevant hence we ignore these two columns. We only require the comments ***(Question)*** and Type of comment ***(Final label)*** to classify the comments. According to the question, ***Final label*** column has been modified:

Requests - request for confirmation, request for information, request for rationale, request for action, request for clarification, request for opinion and action.
Attitudes and Emotions - Criticism, Anger and surprise.

Additional feature considered - Length of the text (number of words in the question)

Classification steps

1. **Data Exploration -** Analyzing the dataset by finding the shape, null values and duplicate values.
2. **Data preprocessing -** We cannot pass the raw data directly to evaluate, they have to be preprocessed (especially for text data). Preprocessing methods that I considered are:
    - Removing special characters
    - Removing Punctuation
    - Removing line breaks
    - Removing space and numbers
    - Removing URLs
    - Removing html characters
    - Removing Patterns
    - Expanding words
    - Lemmatization

Stopwords are not removed though it is one of the important data cleaning process because, in this problem they can give more insights. For example, ‘Can we’ tells us it is a question asked for clarification or a doubt. So after analyzing the dataset, I understood stopwords are crucial for classification, hence it was not removed.

3. **Vectorization -** Text data cannot be passed directly for evaluation, they have to be converted into a numerical data. The process of converting words into numbers is called Vectorization. **Tf-idf** is one of the text vectorization methods, which calculate frequency of a word.
4. **Train-test split -** The data was split in the ratio 90:10 for training and testing respectively.
5. **Evaluation -** The model was evaluated on 3 classification algorithms to see which model gives better accuracy.
    - Logistics Regression
    - Support vector machine
    - Naive bayes

Result and Analysis

	Original dataset		Modified dataset	
	Without added feature	With added feature	Without added feature	With added feature
Logistic Regression	0.74	0.68	0.80	0.78
SVM	0.70	0.56	0.74	0.72
Naïve Bayes	0.72	0.74	0.74	0.74

We can see that **Logistic Regression** performs well overall, and the highest accuracy is **80%** which is from modified dataset and without adding the feature. Modified dataset has better accuracy than the original dataset. And there is not much difference after adding extra feature.

Reference

https://repositorio.ufpe.br/bitstream/123456789/33481/1/TESE%20Felipe%20Ebert.pdf

https://www.w3schools.com/python/python_regex.asp

https://www.geeksforgeeks.org/python-lemmatization-approaches-with-examples/

https://monkeylearn.com/blog/what-is-tf-idf/

https://scikit-learn.org/dev/