Practice VI

Sentiment Analysis

Introduction

- In this practice, the polarity of users' opinions on tourist sites will be determined
- The reviews are about restaurants, hotels and tourist attractions
- Each opinion is rated with a value from 1 to 5, where:
  - 1 is very negative
  - 2 is negative
  - 3 is neutral
  - 4 is positive
  - 5 is very positive
- The aim of the practice is to improve the performance obtained by the basic shared model

Specifications

- Form a team of 3 to 4 people
- The task consists of doing the following
  - Corpus preprocessing
  - Text representation
  - Train Machine Learning model
  - Predict Sentiment Polarity

Corpus preprocessing

- Load the Rest_Mex_2022 corpus
  - *Title* and *opinion* columns must be concatenated and will be used as features
  - *Polarity* column will be used as target (class)
- Apply normalization process
- Create a training set with 80% of the data and a test set with the remaining 20% of the data (set *shuffle=true* and *random_state = 0*)

Text representation

- Create different text representation of the corpus
  - Binarized
  - Frequency
  - TF-IDF
  - Embeddings

Train Machine Learning model

- Split the training set into 5 folds. You can use the *KFold* or *cross_validate* functions
- Train different Machine Learning models tuning parameters (when required) using the 5 folds and calculate the average *f1 macro*
- Select the best adjusted model
- Train the selected model using the full train set

Predict Sentiment Polarity

- Use the trained model to predict the opinions of the test set
- Calculate the average *f1 macro*

Considerations

- It is important to try several normalization and text representation variations. You can lead these experiments based on the results obtained in the previous classification task
- Try different Machine Learning methods and adjust the parameters when required. You can choose the Machine Learning methods based on previous experiences

Additional features

- Features of Vector Space Model can be augmented using linguistic resources. The following are some available resources that can be used:
  - Sentiment lexicons
  - Emojis and emoticons lexicons

Additional preprocessing steps

- Some extra steps can be considered to improve the quality of the corpus
  - Negation handling
  - Word repetition
  - Spell check

Class imbalance handling

- The distribution of classes in the REST-MEX corpus shows a considerable imbalance
- This affects the performance of the models, so balancing methods can be applied to reduce the bias
- Some recommended models are:
  - Over sampling
  - Sub sampling
  - SMOTE

Class imbalance handling

- For the correct application of the balancing methods, the following should be considered:
  - The balancing method should only be applied to the training set
  - Machine learning algorithms should be trained with the balanced version of the training set
  - No balancing method should be applied to the test set and it should retain its original distribution
  - Models trained on the balanced data set will be used to predict polarity in the test set
- Different balancing percentages can be tested to verify the effect that class distributions have on model performance

Evidence

- Source code
- A pdf document with the following sections
  - Introduction explaining the problem to be solved and the corpus features
  - Methodology explaining the steps applied:
    - Corpus preprocessing
    - Text representation
    - Train Machine Learning model
    - Predict Sentiment Polarity
  - Proposed improvements
    - Features augmentation
    - Additional preprocesing steps
    - Others
  - Training results of each experiment
    - Average f score
  - Test results
    - Average f score
    - Classification report
    - Confusion matrix
  - Conclusions and future work