Natural Language

MP2

Classifying Beauty Products' Reviews

Alameda and Tagus

2022

===== Goals =====

Simulate a participation in an evaluation forum (max score: 5)

International evaluation forums are competitions in which participants test their systems in specific tasks and in the same conditions. Thus, training sets are given in advance, and, on a certain predefined date, a test set is released. Then, participants have a short period of time to return the output of their systems, which are evaluated and straightfowardly compared with one another, resulting in a final ranking where the state-of-the-art system is acknowledged. We will simulate such an evaluation forum. A kit-kat will be offered to the winning(s) team(s).

Develop your critical reasoning (max score: 15)

Write a short paper (2 pages) in which you describe your models, present the obtained results and discuss the latter, as well as the given data drawbacks. You should convince me that you have looked at the data and at the returned outputs (not just at scores).

===== Tasks =====

This project is about classifying reviews of beauty products[1]. 5 labels are considered, from the worse classification to the better: =Poor=, =Unsatisfactory=, =Good=, =VeryGood= and =Excellent=. Your task is: a) to build (at least) one model, in python, that classify reviews according with these labels. That is, being given a file with a list of N reviews, your system should return another file with the N predicted labels; b) write a short paper describing your work.

===== Your Models =====

To build your model(s), you can use (alone or combined):

a) a rule-based approach (although by itself results will probably be very poor)

b) a model based on similarity/distances between reviews

d) a machine learning algorithm such as Naïve Bayes, Support Vector Machines, etc. (**deep learning (DL) architectures, neural word embeddings or any kind of neural pre-trained models are not allowed – we will solve the project with DL in the last lab**).

[1] Based on the corpus from https://www.kaggle.com/datasets/skillsmuggler/amazon-ratings

===== Details =====

Groups:

This project should be done in groups of 1 or 2. If you are looking for a colleague to create a group, please add your contact here.

Questions:

As usual, questions should be sent to meic-ln@disciplinas.tecnico.ulisboa.pt (subject: MP2). However, we might release FAQs about the project. If so, please, check them.

Input/output format:

As in an evaluation forum, you will be given a "training" set (train.txt) in which each line has the following format (notice that there is a tab between the label and the review):

label review

Example:
=Poor= *wrong color*
=Good= *good but not great*

A test set (test-short.txt) will be released briefly after the submission of your code and short paper. Each line will have the format:

review

You should run your best model on the test set and return an output file (named results.txt), in which each line has the format:

label

Notice that the line number in which the review appears in the test file should be the same line number of the corresponding label in the results.txt (the evaluation depends on this).

Language and libraries:

You should implement your model in Python 3. You can take advantage of code already available (and I strongly advise you to do so), as long as you identify the source. You can use the following libraries/software: NLTK, Spacy, NumPy, pandas, and scikit-learn. If you really want to use another library, ask first, please.

===== Evaluation =====

Automatic Evaluation (5 points):

- A*ccuracy* will be the evaluation measure.
- If you beat a weak baseline (Jaccard), from now on baseline1, that results in an accuracy of 36,8% (on the test set) you will have 2.5 points.
- If you beat a stronger baseline, based on a Support Vector Classifier and a CountVectorizer, from now on baseline2, that results in an accuracy of 43% you will have extra 2.5 points
- We will randomly select a set of projects and we will run them. If any difference between the reported results.txt and the ones we obtain is found, the group will have a 0 in the project. In order to evaluate your project, we will run the following command:

  python reviews.py –test test.txt –train train.txt > results.txt

Short paper (in Portuguese or English) Evaluation (15 points):

The short paper should be named NUM.pdf (NUM is the number of the group). It should have a maximum of **2 pages**[2], containing the following (**mandatory**):

1. Group ID: The number of the group, and the number and name of each group member should be written in the first two lines.
2. Section "Models and Experimental Setup" (**3 points**): this section contain a clear description of your model(s), all the pre-processing done (if applicable), etc., and, the experimental setup you defined for your experiments.
3. Section "Results" (**1 point**): here, you should present, in a table, along with baseline1 and baseline2 results, your model results, considering accuracy. A confusion matrix will also give you points, as well as the presentation of your results by label.
4. Section "Discussion" (**5 points**): here, I expect you to show me that you have properly analysed the dataset and the obtained outputs (not just by looking at statistics or a confusion matrices). I really want you to look at the sentences that resulted in errors. Explain the most common errors (examples are more than welcome).
5. Section "Future work" (**1 point**): if more time was given to you, explain what you would do to improve your system

Bibliography (if applicable)

We will also score:

- The general quality of your paper (correct syntax, clearness, no typos, illustrative examples, pictures and figures, etc.) (**3 points**).
- The creativity of your approach (**2 points**).

Notice that 3 points will be taken if any instruction is not followed.

[2] If the report has more than 2 pages, we will only evaluate the first two, even if the first one is a cover page with your numbers and names.

===== Submission =====

Part 1 – on November 2th, 2022, **before 1:PM (13h)**, you should deliver, via Fénix (MP2-Part1), a zip file (NOT a rar) containing the project, named after the group number **NUM** (ex: 3.zip).

– the zip file should contain:
  - the file **NUM.pdf** with the short paper
  - a (main) file named **reviews.py** with the project code
  - possible extra python files

The test set, test.txt, will be released on the same day between 1:30PM (13h30) and 1:35PM (13h35).

Part 2 – still on the same day, **between 1:35 PM (13h35) and 11:59 PM (23h59)**, you should deliver, also via Fénix (MP2-Part2), a file named **results.txt** with the results from the given test set, that is, a list of the labels returned by your previously submitted **best model** when it was applied to the given test set. ****NO MANUAL EDITIONS SHOULD BE DONE TO THIS FILE AFTER RUNNING YOUR MODEL; NO CHANGES SHOULD BE DONE IN YOUR CODE OR IN THE OBTAINED OUTPUT. DO NOT ADD THE TEST SET TO THE TRAINING SET WHEN YOU ARE EVALUATING ON THE TEST SET.****

Comments/tips:

- This is not a B.Sc project; this is a M.Sc project: there is a clearly identified problem that you need to solve in the best possible way, but we do not tell you how to do it.
- Remember what you have learned during the class about methodology: try to do a systematic work. Evaluate your models every time you (try to) improve them.
- Pre-processing applied to the training set should also be applied to the test set.
- Attention to blindly removing stop words: your reviews can be empty at the end.
- Understand that language is too complex to deal with each example individually; also remember that your model will need to be able to generalize.
- There is no 100% accuracy (this is a research problem).
- This is a "real" dataset. Datasets in NL have errors and are usually unbalanced (welcome to NLP!). Some categories might be poorly represented (or even nonexistent) (welcome to NLP!). You will also probably find many labels that you don't agree with (welcome to NLP!). You are probably right, but the dataset will not be changed. Write about these situations in your short paper.
- Look at output!!!!!!!!!!!!!!!!! (to predicted labels of the reviews, not just to numbers)

Thank you!

I really hope you enjoy the project and have a good learning experience with it! ♥