Natural Language

MP2

Truthful vs. deceptive hotel reviews

Alameda and Tagus

2023

===== Goals =====

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

International evaluation forums (ex: CLEF, SemEval, etc.) are competitions in which participants test their systems in specific tasks and in the same test sets. 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, although the test set (without the expected output) will be released along with the train set.

Develop your critical reasoning, your communication skills, and your creativity (max score: 15)

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

===== Tasks =====

This project is about distinguish between truthful and deceptive hotel reviews, and to determine their polarity (positive vs. negative). Your task is: a) to build (at least) one model, in Python 3, that classify reviews according with labels TRUTHFULPOSITIVE, TRUTHFULNEGATIVE, DECEPTIVEPOSITIVE, and DECEPTIVENEGATIVE. That is, being given a file with a list of N reviews, your system should return another file with N predicted labels; b) write a short paper describing your work.

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

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:
TRUTHFUL-POSITIVE The sheraton was a wonderful hotel! When me and my mom flew in we were really tired so we decided to take a quick nap. We didnt want to get up! The beds are absolutely to die for. I wanted to take it home with me. The service was great and this was probably one of the biggest if not the biggest hotel ive ever stayed in. They had a really nice restaurant inside with excellent food.

You will also be given a test set (test_just_reviews.txt) in which each line has the format (no labels):

review

During the development of your project, you should create your own test(s) set(s) to evaluate your models. In the paper you should report the results on your own test(s) set(s). However, **for the automatic evaluation of your project**, you should run your best model on the given test set (notice that it has no labels – test_just_reviews.txt) 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 **automatic evaluation** depends on this).

To build your model(s), you can use whatever you want, including taking advantage of code already available (and we strongly advise you to do so), as long as you **identify the source**. The only constraint is: you should implement your model in Python 3.

===== Details =====

Groups:
This project should preferably be done in groups of 2 (cooperation). However, groups of a single person are also allowed. If you are looking for a colleague to create a group, please check the sheet we will make available on Fénix.

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

===== Evaluation =====

Automatic Evaluation (5 points):

- A*ccuracy* will be the evaluation measure.
- If you beat a weak baseline (Jaccard) that results in an accuracy of 58.5% (on test_just_reviews.txt) you will have 2.5 points.
- If you beat a stronger baseline, based on a Support Vector Classifier and a tf-idf that results in an accuracy of 88.0% (on test_just_reviews.txt) you will have extra 2.5 points.

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**[1], 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 lines.
2. Section "Models" (**3 points**): this section contains a clear description of your model(s) and all the pre-processing done (if applicable).
3. Experimental Setup and Results (**2 points**): in this section you should detail your experimental setup (evaluation measures, datasets used, dev set, parameters, etc.) and present your model(s) results. These should be presented in a table, considering accuracy (general results, and also the results obtained by label), as well as a confusion matrix. **As previously said, the expected labels of the given test set are not provided (test_just_reviews.txt), so you should create your own test(s) set(s) from the training set and report the results on your own test(s) set(s). The given test set (test_just_reviews.txt) should only be used to generate the file results.txt.**
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 mandatory).
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)

[1] 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.

We will also score:

- The general quality of your paper (correct syntax, clearness, zero typos, illustrative examples, pictures and figures, etc.) (**2 points**).
- The creativity of your approach (**2 points**).

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

===== Submission =====

On October 27th, 2023, you should deliver, until **11:59 PM (23h59)**, via Fénix, 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
– a file named **results.txt** with the results from the given test set, that is, a list of the labels returned by your **best model** when it was applied to the given test set (test_just_reviews.txt).

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 sometimes unbalanced (this one is not). You will also probably find many labels that you don't agree with. You are probably right, but the datasets will not be changed. Discuss these situations in your short paper.
- Look at the output!!!!!!!!!!!!!!!!!! (not just to numbers)

Thank you!

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