CE 314/887 Assignment 2

Text classification

December 2022

Deadline: Jan 9th 2023

Build a text classifier on the IMDB sentiment classification dataset, you can use any classification method, but you must training your model on the first 40000 instances and testing your model on the last 10000 instances. The IMDB dataset will be uploaded on the moodle page for you to download.

Your code should include:

1: Read the file, incorporate the instances into the training set and testing set.

2: Pre-processing the text, you can choose whether you need stemming, removing stop words, removing non-alphabetical words. (Not all classification models need this step, it is OK if you think your model can perform better without this step, and you can give some justification in the report.)

3: Analysing the feature of the training set, report the linguistic features of the training dataset.

4: Build a text classification model, train your model on the training set and test your model on the test set.

5: Summarize the performance of your model (You can gain additional marks if you have some graph visualization).

6: (Optional) You can speculate how you can improve your works based on your proposed model.

After you build such a model and test on the test set, you should write a report (no longer than three pages in A4, with Arial 11 fonts) to summarize your work.

(You can use the existing algorithms on github or kaggle, but you must not directly copy and paste their code!

However, you are not allowed to use the Naïve Bayes algorithm and VADER classifier, which practiced in Lab 4)

Suggestion: some bonus points:

Have necessary comments on your code

Have proper reference on your report

Have graph visualization on your report

Investigate more evaluation methods, like not only show the P R F score, but also run multiple times and show the standard derivation on P R F (I am sure you can find more evaluation methods.)

Write your report like a mini-conference paper (you can learn from this paper:

- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

)