Determining tweet relevance for a given subject from the analysis of its content

1 Description

Internet can be used as one important source of information for machine learning algorithms. In particular, twitter has become a valuable tool of information for companies and social agents. Supervised classification methods are used to identify and classify the reaction to products and events [1, 2, 3, 4, 5].

2 Objectives

The goal of the project is to predict what is the relevance of a tweet regarding self-driving. This can be posed as a classification problem in which the tweets can be classified in two classes: relevant or not relevant. A database of tweets will be used for the analysis [1]. A supervised classifier should be created to predict the sentiment contained in each tweet. An added difficulty is that this is an unbalanced problem in which only few of the tweets can be classified as "not relevant". Therefore, the AUC metric or average class classification are recommended over the use of the accuracy.

The student should: 1) Design any preprocessing of the tweets in the dataset (see suggestions below); 2) Define and learn the classifier using the training data. 3) Design the validation method to evaluate the accuracy of the proposed classification approach.

As in other projects, a report should describe the characteristics of the design, implementation, and results. A Jupyter notebook should include calls to the implemented function that illustrate the way it works.

3 Suggestions

- Formalize the task as a binary classification problem.
- Use the `pattern` Python package `https://github.com/clips/pattern` to extract features from the tweets.
- See example `https://github.com/clips/pattern/blob/master/examples/03-en/07-sentiment.py` to see tools for sentiment analysis in `pattern`.
- Implementations can use any other Python library.
- If classes are not well balanced you may use performance measures different to the accuracy.

References

[1] Nicholas Beauchamp. Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, 61(2):490–503, 2017.

[1] The "Twitter sentiment analysis: Self-driving cars" dataset can be downloaded from `https://www.crowdflower.com/data-for-everyone/`.

[2] Aron Culotta. Training a text classifier with a single word using twitter lists and domain adaptation. *Social Network Analysis and Mining*, 6(1):8, 2016.

[3] Jimmy Lin and Alek Kolcz. Large-scale machine learning at twitter. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 793–804. ACM, 2012.

[4] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, volume 11, pages 281–288, 2011.

[5] Agus Sulistya, Abhishek Sharma, and David Lo. Spiteful, one-off, and kind: Predicting customer feedback behavior on twitter. In *International Conference on Social Informatics*, pages 368–381. Springer, 2016.