Inferring the actitude toward self-driving cars from tweets using supervised classification algorithms

1 Description

Internet can be used as one important source of information for machine learning algorithms. In particular, twitter has become a valuable source 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 infer the reactions of users to self-driving cars from their tweets. The problem can be transformed in classifying the tweets in different classes according to the sentiment they express about self-driving cars. 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.

The student should: 1) Design any preprocessing of the web pages 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 classification problem where the classes of the problems are: very positive, slightly positive, neutral, slightly negative, or very negative.
- 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.