Final Project

L545/B659

Due: Thursday, December 16, 2021

You will work in groups of 2-3 people. Please send me an email with your groups by October 30. Make sure that every group has one person with programming experience. If you are not part of a group by then, I will assign you to a group.

1. Read this paper: `https://aclanthology.org/S16-1003.pdf`

2. Part A
   - Extract a bag-of-words list of nouns, adjectives, and verbs for all targets individually. Then create feature vectors for all training and test data (separately) for all targets.
   - Perform classification using Support Vector Machines (SVM) and default settings.
   - Can you improve the results when you optimize the settings?

3. Part B:
   - Then extend your data set to include features using the MPQA Subjectivity lexicon (`http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`). Decide on a good way of using this information in features. Explain your reasoning. How do the results change?
   - Can you use the Arguing Lexicon (`http://mpqa.cs.pitt.edu/lexicons/arg_lexicon/`)? Do you find occurrences of the listed expressions? How do you convert the information into features? How do these features affect classification results?

4. Part C:
   - Parse your training and test data using MALTparser and the predefined model. Then extract dependency triples form the data (word, head, label) and use those as features for the stance detection task instead of the bag-of-words model. How does that affect the results?

5. Part D:
   - What happens if you use all words ONLY as features?
   - What happens when you use bi-grams along with unigrams as features?
   - What happens when you use uni-grams, bigrams and trigrams as features?
   - Note: You will have to use TF-IDF vectorizer for feature selection in this part

6. Write up your findings in a short paper (one paper per group) and submit via canvas. Make sure that you include all your results, preferably in tables, and discuss these results in the text. Describe the experiments in enough detail that I could replicate them without problems. And make sure that the paper answers all the questions above. Note that the readability of the paper will have consist of a good part of your score for the project.

7. Submit your code, copies of the system output along with your paper.

8. Extra challenge for extra credit: If you want an additional challenge, repeat all experiments using the random forest implementation in scikit learn, optimize the parameters and compare the best implementation of SVM with it for each of the parts. (**20 extra points**)

Start early with the tasks so that you can ask questions when you get stuck. If you have any questions or run into any problems that you cannot answer within your group in a reasonable amount of time, talk to me!!! I know that this is a rather challenging project in many respects. But don't forget: "CL is a social science – alone, it's too hard." (adaptation of a quote from one of my favorite professors, Uwe Moennich)