Assignment P2

Part 1 (20 points)

Part 1 of the assignment involves annotating a sample dataset (*shared with you through google drive, check your ncsu email and google drive*). You are required to label a sample dataset containing around 120 pairs of sentences based on the Annotation instructions.

Annotating a dataset is an essential part for any classification task and since natural languages allow us to express the same information in so many ways makes this task all the more interesting and sometimes challenging. Do not slack on this task as this will impact part 2 of this assignment that makes use of the annotations you provide.

P1 is due on 5:00 pm ET 10/24/21

Part 2 (80 points)

The annotated response types are aggregated and released to you for the Phase II task.

Task:

- Extract the baseline features
- Propose two additional features that you think would contribute to type classification
- Vectorize the features and classify response types
- Evaluate classification performance
- Write a report. Include a link to the lexicon in the report if any is proposed as additional features

Baseline features (Include these features as your baseline model):

- word/sentence embeddings

Some ideas for additional features(be creative):

- The number and types of pronouns, such as “I”
- The length of the sentences
- Rhetorical questions
- sentiment analysis of comments and responses.
- Part of speech

One challenge of this task is that you need to find a way to “combine” the vectorized representation of the challenger's comment and OP’s response. It’s also your decision whether you want to use the story description as additional input.

For example, to use word/sentence embeddings, a straight-forward way is to learn a sentence embedding of a comment and a response, concatenate the two vectors, and use a simple classifier such as SVM. For high-level features such as sentiment scores, be creative on how to incorporate the scores for a comment and its response.

Model:

You should experiment with multiple classification models (such as SVM or neural-network-based classifiers) and choose the best to report results for. You are required to report the results for the following feature sets:

- set 1: only baseline features
- set 2: baseline features + two additional features (that you proposed)

Evaluation:

- accuracy
- precision, recall, and F1 score

We don't have a performance benchmark. However, your report should reflect the effort and thoughts you put into improving the performance. Based on the performance of the model on the two feature sets, analyze why your proposed additional features **contribute or not** to type classification.

Deliverables for Part II

Upload a single zip file (make sure not to include any unnecessary superfluous files that may inflate the file size beyond the submission locker's limit) to the submission locker. Include the following in the zip file:

- A report that:
    - Describe your baseline solution and proposed solution in detail. Organize your report into different sections such as data source(if you use extra data), preprocessing, tokenization, modeling building, modeling training, modeling evaluation, results and so on.
    - Compare the performances of the baseline and two additional features.
    - Also describe how you train and evaluate your model, e.g., how you split the training set.

- A jupyter notebook that:
    - Self-contained. We will download your notebook and run in colab. You can assume the data is in the current directory(no need to manipulate the paths). You need to test your code before you submit, and make sure the dependencies are

properly installed at the beginning cell. E.g. !pip install Spacy(in colab, ! means shell commands)

  - Well-documented but not too verbose. Organize your notebook into sections such as data processing, text tokenization, model building and so on. Also make comments with your code to explain what your code does. You should not expect the instructor to speculate on your code.
  - Show some key intermediate execution outcomes. When you save your colab notebook, the outcome of each cell is also saved automatically. You can remove some trivial outcomes such as installation information as you see fit.

P2 is due on 5:00 pm ET 11/07/21