PROJECT

DIVISION OF PROBLEM STATEMENT:

1. Task 1 : Classify sentence into subjective or objective class
2. Task 2 Part A : For subjective sentences obtained from Task1, Identify words that are introducing subjectivity in the sentence.
3. Task 2 Part B : Suggest alternative words for subjective words so that sentence will have no or subtle subjectivity but have same/similar meaning as the original one.

APPROACH:

INPUT:

Representation of words in source sentence:
1. Word2Vec or Glove embedding
2. BERT embedding (If possible)

Representation of source sentence:
1. Average of embeddings of the words in the sentence
2. Concatenation of embeddings of the words in the sentence by limiting length of the sentence. (To decide length threshold, we may need to analyse data to check length of sentences)

BASELINE MODELS

Baseline models for Task1 : (Input: Average embedding representation of sentence)
1. Random Forest
2. Naive Bayes
3. Logistic Regression
4. SVM
5. Neural Network

Baseline model for Task 2 Part A: (Input: Subjective sentences identified in Task 1)
1. Identify subjective words by checking if any word is present in the lexicon collection of lexicons
2. Identification of subjective words by using sentiment analyser like Stanford's coreNLP tool, Textblob, Affinn, etc. and explore if there are any functions which can give us subjective/ sentiment score for words in the sentence or give subjective words in the sentence.
3. Identification of subjective words by using WordNet to find sentiment of words in the sentence
4. Explore more ways/tools to identify subjective words in the sentence. Also, check if it is possible to provide subjectivity score for words if multiple subjective words are found in one sentence for above models.

Baseline model for Task 2 Part B : (Input: Subjective sentences identified in Task 1 with words introducing subjectivity in those sentences obtained in Task 2 Part 1)
1. For a subjective sentence, for each subjective word in the sentence (in the order of decreasing subjectivity scores):
    1. Step 1 : Get similar words or synonyms
        1. Use Wordnet which returns list of synonyms called Synset
        2. Use embeddings to get similar words using cosine similarity.
    2. Step 2 : Get subjectivity score
        1. For each word from the list of similar words obtained in step 1, get subjectivity score of the word using any of the model used in Task 2 Part A.
    3. Order above words in decreasing order their subjectivity score and return top N words as alternative.
2. If a subjective sentence has more than one subjective words, then we can either suggest alternative words for most subjective word only or we can suggest alternative words for every subjective word but in order of decreasing subjective scores i.e. user should consider replacing most subjective word with alternative word first.

ACTUAL MODELS:

Actual Model for Task 1:

1. LSTM (Input: Concatenated embedding representation of sentence (Use of linguistic features If possible)
2. Detector Model mentioned in the paper (Use of linguistic features If possible)

Actual Model for Task 2 : (Part 1 and Part 2 combined) :

1. Model mentioned in the paper: Encoder-Decoder (Bi-LSTM encoder, LSTM decoder with Attention mechanism and denoising objective function)

Pipeline (Task 1 —> Task 2):

- Use only WikiEdits dataset to train models in Task2.
- Subjective sentences identified by models in the Task1 which do not belong to WikiEdits dataset will be given to models trained for Task2 as unlabelled/test data which will identify subjective words and suggest alternatives. (Basically, data other than WikiEdits will be rated as test data for models trained for Task2)