Assignment 1: Sentiment Analysis Text Classification (15 points)

CS 410/510 Natural Language Processing, Fall 2023

In this assignment, you will take your first steps into the fascinating field of natural language processing (NLP) by exploring text classification, a fundamental NLP task. The goal is to build a text classifier that will predict whether a piece of text is "positive" or"negative". Your task is to see how accurate a classifier you can build, depending on the machine learning approach and on the various features you will be asked to implement.

1 Instructions

1. **Data Preparation**: You are provided with a dataset containing customer reviews of products. The dataset is the "Multilingual Amazon Reviews Corpus" (1) in json format. It includes several columns. For this assignment, we will be using a small subset of the original dataset and are concerned with the following two columns:

   (a) "review_title" (the title of the review), and

   (b) "stars" (an integer 1 or 5 indicating the number of stars, which is considered as the sentiment label, with 1 indicating "negative" and 5 indicating "positive").

   Warning: *As is the case with most 'natural language' text, this data was collected from public websites and is mostly unfiltered. Therefore, it is possible that some text may be disturbing or you may not agree with it.*

   - The dataset is pre-split into a training set and a test set.
   - Load the dataset using Python and the appropriate library (e.g., pandas).
   - (optional) Perform any necessary data preprocessing steps, such as lowercasing, tokenization, and removing punctuation.

2. **Feature Engineering** An important part of text classification is choosing the set of features and their representation that would help you build the most accurate classifier. Some of the features that you have already seen in class include n-grams. As part of this assignment, you will further experiment with a variety of features to find the best performing features for your models. As an example, you can use any of the feature types listed below or use additional features you come up with yourself.

   - Num words: the number of words per title

- *N*-grams: unigrams, bigrams, and trigrams
- Cue words: the initial unigram, bigram, and trigram in a post
- Repeated punctuation: features to capture punctuation such as ??, ??????, !!!, or ?!
- Part-of-speech tags: features to count the number of nouns, verbs and adjectives in the text (count_noun, count_verb, and count_adj)
- Any sentiment lexicon (e.g., NLTK Vader[1])

3. **Text Classification Model**: Now, it's time to build a text classification model. Implement a simple text classification model using any machine learning library of your choice (e.g., scikit-learn or NLTK). Follow these steps:
   - Choose *two* suitable algorithms for text classification (e.g., Naive Bayes, Logistic Regression, or a basic neural network).
   - Vectorize the text data (convert text into numerical features).
   - Train the model using the training set.
   - Evaluate the model's performance on the test set using appropriate evaluation metrics (e.g., precision, recall, F1-score).
4. **Results and Analysis**: Provide a detailed analysis of your model's performance by comparing the output of the two algorithms. Include the following:
   - F1-score and other relevant metrics.
   - Confusion matrix.
   - Any observed challenges or limitations in the classification task.
   - Suggestions for improving the model's performance.

2 Submission Guidelines and Grading

Submit a PDF of your Colab notebook on Canvas, under "Assignment 1". Include any additional files or resources used. Ensure your code is well-documented and organized.

To convert your Colab notebook to PDF format, you'll need to install a Python package called nbconvert, a command-line tool that allows you to convert Jupyter notebooks to various formats. To install nbconvert, open a new code cell in your Colab notebook and run the following command:

```
!pip install nbconvert
```

Once nbconvert is installed, you can convert your Colab notebook to PDF format using the following command (replace <notebook-name> with the name of your Colab notebook):

```
!jupyter nbconvert —to pdf <notebook–name>.ipynb
```

[1] `https://www.nltk.org/howto/sentiment.html`

Your assignment will be assessed based on the following criteria:

- data preparation (3 points),
- implementation of the text classification model (5 points),
- model evaluation and analysis (5 points),
- and overall clarity and organization of the assignment (2 points).

Good luck with your first text classification assignment!

References

[1] Keung, Phillip, Yichao Lu, György Szarvas, and Noah A. Smith. "The multilingual amazon reviews corpus." arXiv preprint arXiv:2010.02573 (2020).