Ethics for Natural Language Processing 2024 Homework 0

TECHNISCHE UNIVERSITÄT DARMSTADT

Due until Wednesday, 08.05.2024 at 11:59pm

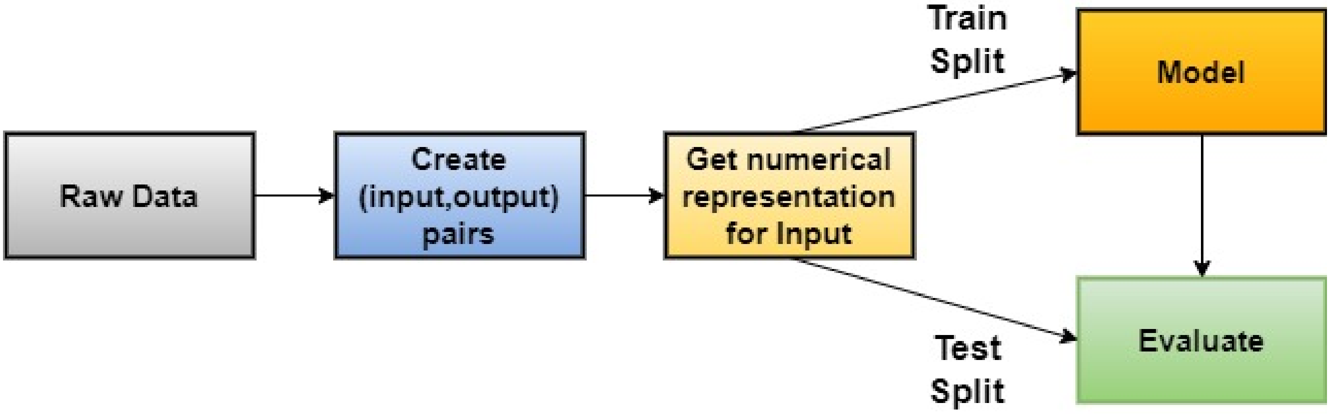
Submission Guidelines for Homework

- This homework is worth 20 points
- Use the .ipynb file as a template.
- Submit your code and answers in a single .ipynb notebook.
- Extra credit shall be given to well-structured submissions.
- In case of questions or remarks, please contact:
  - Aishik Mandal, aishik.mandal@tu-darmstadt.de

This is a warm-up homework to get you familiar with basic NLP operations. Before you start, make sure you read the Submission Guidelines instructions associated with this homework for important setup and submission information. Additionally, we encourage you to use the notebook provided with this homework as a template: we have already put a lot of code in it and it will give you a head-start on the assignment.

1 Text Classification

We start with text classification – a classic NLP task: given a text, assign it a label. There are many examples of this task, from sentiment analysis to natural language inference to assessing grammatical correctness. Here, we will predict the ratings of a product from its review text. We will use the Amazon Product review dataset for this purpose. This task belongs to the broad category of emotion recognition.


Train Split
Model
Raw Data
Create (input,output) pairs
Get numerical representation for Input
Evaluate
Test Split

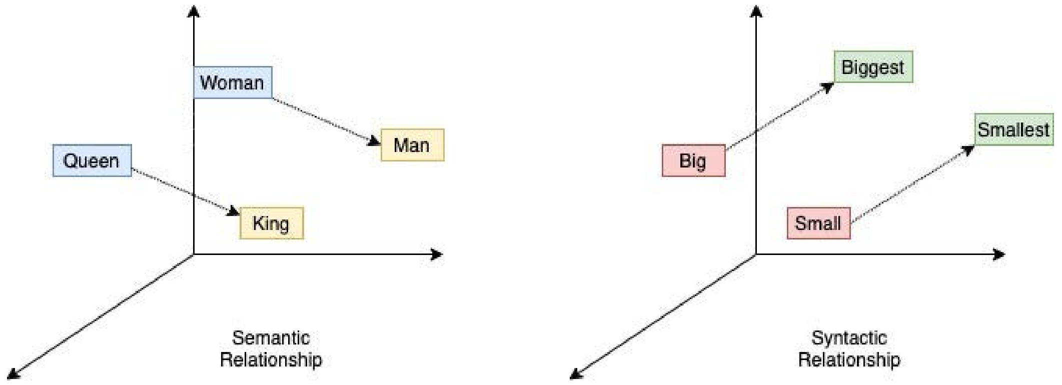
1.1 Loading Dataset (2 Points)

Modern NLP datasets commonly look like a csv or json file, that contains the input text and the output label. In this case, we have the reviews as inputs and the ratings as outputs. The Amazon Product Reviews dataset consists of various categories of products. In this problem, we provide two small splits, the "All Beauty" split and the "Gift Cards" split. You are given three different csv files for each split.

The three different csv files correspond to training, validation and testing data. Your first task is to use these csv files to construct a dataset object consisting of (input, output) pairs. For this task, you will mainly be using the Pandas library in Python.

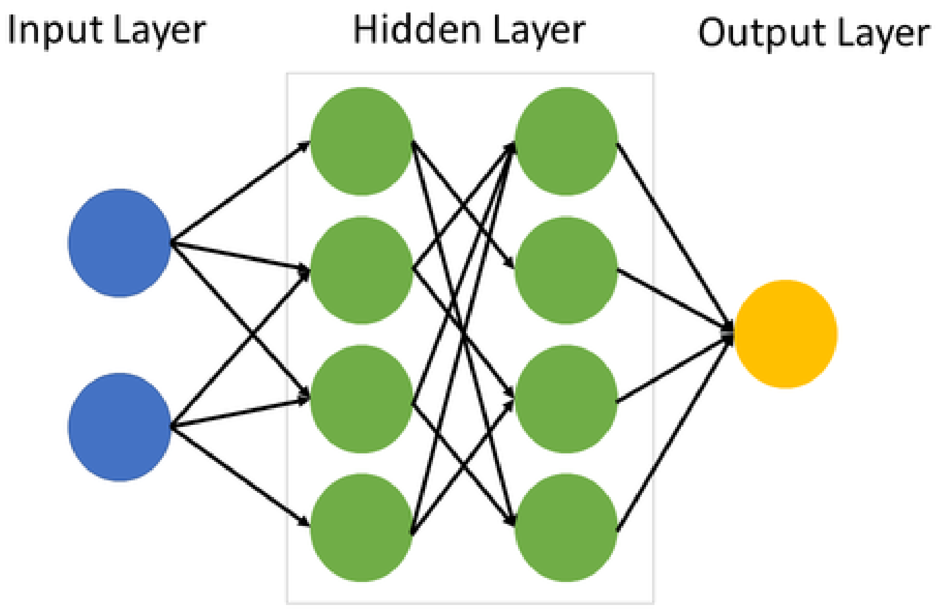
1.2 Word Embeddings (2 Points)

Next, we need to represent the text numerically. There are several ways to do this. Here we will use a classic approach: a tokenizer followed by a static word embedding. A tokenizer breaks natural language text into chunks of information that can be considered as discrete elements, for example, words. A static word embedding is a table where each word is associated with a vector. You will use the well-known word2vec embeddings to get vector representation for each word. There will be some words which are not in the vocabulary of word2vec. For simplicity, you will simply skip these words for now. As result, you will have a list of vectors – one for each word. To get a representation of the whole product review, you will get the average of the vector representations of each word in the review apart from the words that are missing from word2vec. This is called mean pooling, a technique often used for dimensionality reduction.



1.3 Neural Network (3 Points)

You have represented your product review as a vector – this is your input. Now you can build a mapping to map the input to the desired output – the label. This is a non-trivial mapping, because the review is now represented as a dense vector, and the relationship between individual vector values and the label is non-linear. You will be using a multi-layer neural network to approach this. You will use the training split from the dataset to train a model.



Artificial Neural Networks

1.4 Result Analysis (3 Points)

Finally, you evaluate your trained model on the validation and test split of your dataset. Plot the confusion matrix of your predicted vs the actual outputs. Now, you might notice that the accuracy seems good. But in fact, the results are not that great. Why? From the confusion matrix, identify the problems with the results.

2 Ethical Considerations

2.1 Adversarial assessment (6 Points)

We have built and evaluated a classifier for emotion recognition in product reviews. Assume that we now build an exactly same classifier, but this time it predicts a grade based on a student's essay. We take text as input, and predict a grade from 1 (great) to 5 (bad). Our hypothetical classifier works remarkably well: 99% accuracy[1]. Assess this classifier adversarially. (1) Who would benefit from such classifier? (2) Who can be harmed? To get the data, we asked a hypothetical colleague from the computer science department to give us essays and grades from a last years' course. (3) Is the data representative? (4) Is the data collected ethically? (5) Is the modeling approach (neural network) appropriate for the task? (6) Can you think of any dual use for such classifier?

2.2 Ethics board review (4 points)

Below are two research project ideas. In your opinion, which one (or maybe both?) would require a review by an ethics board / IRB before it can start, and why?

- Project A. We want to build a tool that helps researchers find related works in research papers. The tool will analyze the content of a research paper while the user reads it, and will use a large external database to find other papers related to each sentence – even if they are not cited in the

[1] No real essay evaluation system will perform that well. Even two human teachers are unlikely to agree with each other in 99% of cases.

main paper. The results will be shown to the user as highlights on the text. To build the tool, we will conduct a formative study with ten participants, which will include an interview, a short survey, and a screen recording of the users' interactions with a tool prototype. We will then develop the tool by using publicly available research papers. Once the tool is developed, we will ask 50 more participants to use it and rate its helpfulness on a standard questionnaire. We will also record the participants using the tool via screen recording, and analyze the recordings to improve the usability of the tool.

- Project B. The computer science department will conduct an experiment to find out how the politeness of a chatbot like ChatGPT affects the way people interact with it. 100 participants will be recruited online for a study in which they will have to collaborate with another person over a chat window to solve a simple task. The chatbot will be deployed in three different "politeness" settings: very polite, neutral and rude. To ensure natural behavior, the participants will be told that the study is conducted by the psychology department to study collaborative problem-solving, and will not be made aware that they are talking to a chatbot. Collected data will be anonymized to not contain personal information, and will be analyzed in terms of the number of steps required to solve the task, as well as text content of people's interactions with the chatbot.