# Ethics in Natural Language Processing 2024 Homework 1

TECHNISCHE UNIVERSITÄT DARMSTADT

Due until Wednesday, 29.05.2024 at 11:59pm

**Submission Guidelines for Homework**

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

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, as we have already put a lot of code in it. Also, this will give you a head start on the assignment.

## 1 Word Vector Visualisation

In the last homework, we were introduced to word vectors that are used to represent a text numerically. Now, we will try to visualise them by plotting some word vectors in the 2D plane and observe their relative positions. For this, we will use word2vec vectors(also known as word embeddings).

### 1.1 Dimension Reduction(2 points)

As you know, word2vec gives 300-D word embeddings. However, to visualise them on a 2D plane, you will need to perform a dimensionality reduction operation. Follow the following steps to perform it:

- Put the first 10000 word2vec embeddings in a matrix. We only use the first 10000 words to reduce memory usage. Otherwise, you will get an out-of-memory error.
- Use TSNE to perform a dimensionality reduction operation from 300D to 2D

### 1.2 Visualising word2vec vectors(2 points)

Plot the 2D word2vec embeddings obtained for the following words:
['Football', 'Hockey', 'Baseball', 'Tennis', 'Field', 'Court', 'Law', 'Science', 'Literature', 'Computer', 'Games']

## 2 Bias in Word Vectors

Next, you will analyse the gender bias present in these word vectors and the bias they introduce when they are used for training a model.

## 2.1 Simple Bias Analysis(4 points)

- Find which words are most similar to “job” and “women” and most dissimilar to “men”.
- Find which words are most similar to “job” and “men” and most dissimilar to “women”.
- Comment on the bias you observed through this example.
- What kind of bias among label bias, selection bias, overamplification bias and semantic bias did you observe here?

## 2.2 Bias Analysis in models trained with word2vec embeddings

Now, you will analyse biases formed in models trained using word2vec embeddings. For this, you will use the SemEval-2018 Task 1: Affect in Tweets. You will use the train and test split of the English valence regression dataset. The dataset contains three columns:

- Tweet - The tweet itself as a string, the input.
- Intensity Score - The sentiment’s valence of the tweet in the range [0, 1], the output.
- Affect Dimension - You can ignore it. It is ’valence’ for all of the data points.

### 2.2.1 Model training(4 points)

You will follow the standard pipeline for training a logistic regression model:

- Preprocessing (e.g., removing stopwords and punctuation, tokenization)
- Transforming the tweets’ tokens into a single 300-dimensional vector and then performing mean-pooling to get 300-D vector representing the tweet. You will also convert the valence intensity score to a binary valence label, i.e. 0 or 1. For this, you can use a threshold of 0.5. If the intensity score is less than or equal to the threshold, set it to 0, and if it is greater, set it to 1.
- Applying logistic regression to predict the valence label. You can use the logistic regression from scikit-learn library.

Once you have trained the classifier, report its accuracy on the train and test set of the given dataset.

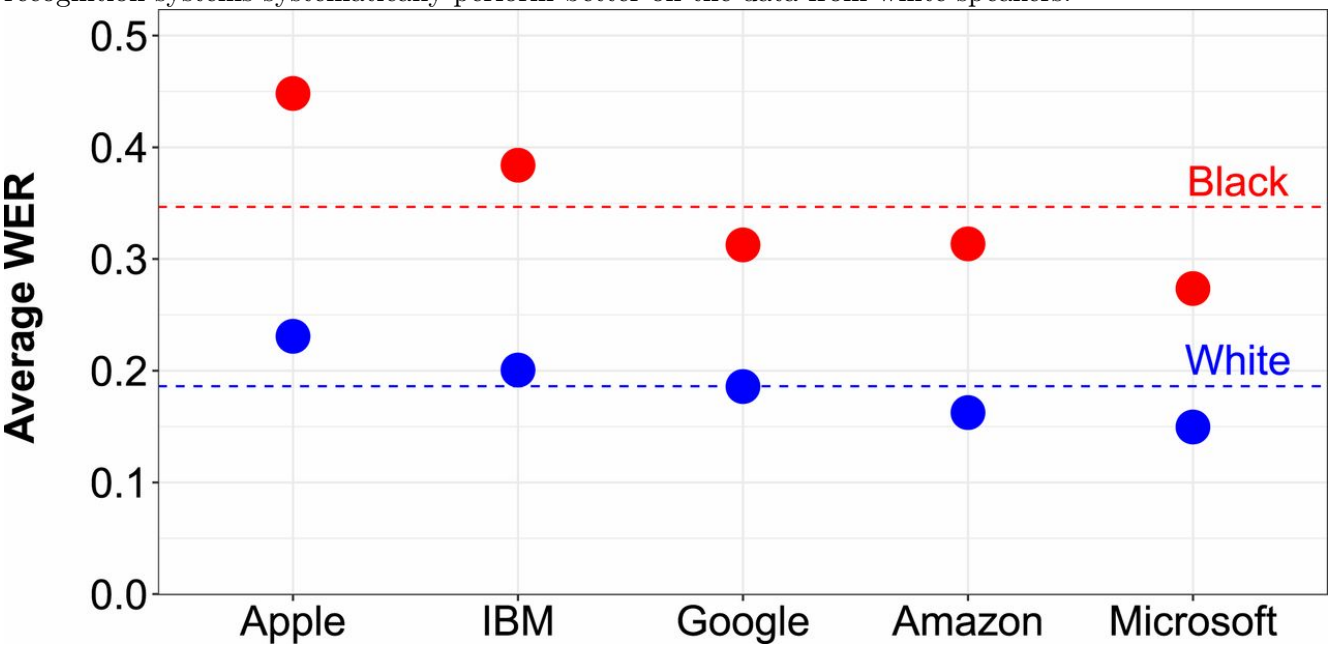
### 2.2.2 Evaluating Gender Bias in Downstream Tasks(4 points)

You will evaluate your trained model from the previous step on the Equity Evaluation Corpus (EEC). It consists of 8,640 English sentences carefully chosen to tease out biases towards certain races and genders. In this corpus, every sentence is built using two elements: Person and Emotion Word. Each sentence in the corpus belongs to the template <person> feels <emotion word>. Thanks to this systemic construction from templates, the sentences are paired by gender, i.e. the EEC data is built of pairs of sentences that are all the same except for a gender noun. So, ideally, if you divide the dataset into male and female subsets, you should have an identical probability distribution for valence labels.
Let us see if that is the case. Your tasks to test this are as follows:

- Divide the dataset into male and female subsets (already done in the template).
- From each, pick subsets related to a certain emotion. In this case, use “joy”.
- Plot the probability distribution for the valence label 1 for emotion in the male and female subsets. For this, you can use distplot from the seaborn library.
- Do you observe any differences in the distribution plot among the two genders? If so, which gender is more likely to be predicted as joyous? (Note that the difference between the distribution will be very slight. Don’t expect a huge difference in the distribution plots.)

## 3 Understanding type of bias(4 points)

As explained in the lecture, ML models can exhibit biased behaviour in two ways: outcome disparities and error disparity. This can happen due to five sources of bias in ML models: label bias, selection bias, overamplification bias, semantic bias and design bias. Let's look at this study: Racial disparities in automated speech recognition. The study compared several speech recognition systems in terms of their Word Error Rate (WER) for African-American vs white speakers. As the plot demonstrates, the speech recognition systems systematically perform better on the data from white speakers.




- What kind of biased behaviour do we observe here? Is it outcome disparity or error disparity?
- The observed behaviour is probably a result of multiple bias sources. While we don't know how exactly the speech recognition models from Apple, IBM, Google etc. are built, we can safely assume that these models are trained on large amounts of speech-to-text data: audio signal paired with text transcription. This data might not necessarily represent the population that is using the system. Which of the five bias sources could be at play here? Provide a brief explanation for your answer.