Problem Statement -

Textual similarity is an important problem in NLP. The broad goal of textual similarity is to measure the extent of similarity between a given pair of text fragment (sentence/paragraph etc) based on a specific aspect/criterion (such as topic, sentiment, etc).

In this problem, you will be given a set of pairs of text and the end goal is to determine if they are similar or dissimilar based on some form of **gender bias present in the text**. The forms of the gender bias could be: i) firstness: where a gender (male/female) is always mentioned first, ii) stereotyping of a particular gender, iii) subordination: where the text reflects a gender is subordinate compared to other.

Please refer Here to get the dataset and details about it.

Training data format: The training data will contain a set of text pairs (p1 p2) along with their labels (0 or 1), where 0 indicates p1 and p2 can be both biased or both unbiased, similarly 1 indicates if one is biased but the other is unbiased. There are two files for training data. The first file (name: text-and-id) contains the text (2 nd column) and its unique id (1 st column) in each line, while the second file (name: pairs-label-training) contains the ids of the two text fragments and the corresponding label (0/1) in each line.

Desired output: Given a text pair (p1 p2), your goal would be to mark this as 0 or 1. The meaning of 0 and 1 is the same as in the training data.

Test data: Test data contains a set of text pairs and you have to produce the 0/1 tag for each pair in the test data.