Paper Dissection: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings *(Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.)*

As fourth assignment of the ML4NLP course we were given the choice between giving a small presentation or dissecting a paper and our group chose the latter. This paper dissection follows the questions of slide four (from the slide deck for assignment four). Each question is answered in a small subsection. The subsections are kept concise and aim to answer the questions adequately.

What is it about? What problem does it try to solve? Why is it interesting?

Bolukbasi et al. (2016) paper is provocatively titled "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". As the title of the paper hints at, Bolukbasi et al. (2016) first present the problem of word embeddings reflecting real world bias. They show that even word embeddings trained on Google News articles exhibit serious sexist (male/female) gender stereotypes. They prove that gender bias is an issue by capturing the direction in the word embeddings and further show that neutral words are linearly separable from gender definition words using word embeddings. While this behavior can be expected if the embeddings are trained on old literature or other texts reflecting gender stereotypes, we would have expected news articles to make use of fairly "neutral language". Keeping in mind that machine learning algorithms are used for tasks that seriously impact people's lives, such as risk assessment tools which help judges determine sentence length and probation option or hiring algorithms used to help HR-departments, the gravity of any sort of unwanted and discriminatory bias becomes evident.

They go on to present two methods they developed, named "hard-debiasing" and "soft-debiasing", for debiasing word-embeddings. They make use of the linear separability to modify embeddings and remove gender stereotypes such as receptionist and female, while maintaining desired associations such as between the words queen and female. They finish by empirically showing that the priorly mentioned methods hold up to their claims, meaning that their embeddings can be used in application without amplifying any gender bias.

Which ML methods are used? What is the main innovation of the paper?

As the title says the main ML method that is used throughout the paper are word embeddings. The embeddings used in this paper are based on the w2vNEWS embeddings. To be more specific they consider the 50,000 most frequent words (only lower-case words and phrases consisting of fewer than 20 lower-case characters are considered, while words with upper-case letters, digits, or punctuation were discarded). These embeddings were normalized to unit length. The paper's main innovations are the two debiasing method they present. Firstly, they present a function that completely equalizes sets of words outside the subspace and thereby enforces the property that any neutral word is equidistant to all words in each equality set. The disadvantage of their hard-debiasing function is that it removes certain distinctions that are valuable. For example, one may wish a language model to assign a higher probability to the phrase to grandfather a regulation, than to grandmother a regulation as the first has a meaning, while "to grandmother something or someone" does not – equalizing the two removes this distinction. For cases like this they present their second function, a soft-debiasing function, that reduces the differences between these sets while maintaining as much similarity to the original embedding as possible, by using a parameter that controls the trade-off.

What are the take aways? I What are possible problems of the approach? Think critically!

The key take-aways are the importance of taking bias and discriminatory tendencies into account when working with ML and designing ML methods and algorithms. A problem of this approach is that it only debiases gender-specific bias, however it is safe to assume that there are other demographics (such as race) that this also applies to. This must be kept in mind when working with ML methods.

What does one need to know for understanding the paper? Add the resources that were helpful for you.

One needs to understand the principle of word-embeddings (https://www.youtube.com/watch?v=mWvnlVw_LiY, https://www.youtube.com/watch?v=BWaHLmG1lak) and one must have basic knowledge of linear dependence (https://www.khanacademy.org/math/linear-algebra/vectors-and-spaces/linear-independence/v/linear-algebra-introduction-to-linear-independence) and algebra. Those sources should be enough to gain a basic understanding of how the research was conducted and what steps they took.