Exploring Bias in German and Dutch Natural Language Processing Models

Degree programme: Master of Science in Engineering
Thesis advisor: Prof. Dr. Mascha Kurpicz-Briki
Expert: Dr. Elena Nazarenko

This thesis investigates biases in German, Dutch, and multilingual natural language processing (NLP) models. It examines models commonly used in research as well as state-of-the-art models from OpenAI and voyageAI. The data used for the metrics is partly derived from workshops aimed at detecting language-specific biases in the labor market. The results show that the identified biases are reflected in almost all multilingual models, most German models, and some Dutch models.

Leander Rankwiler
Data Science

Introduction

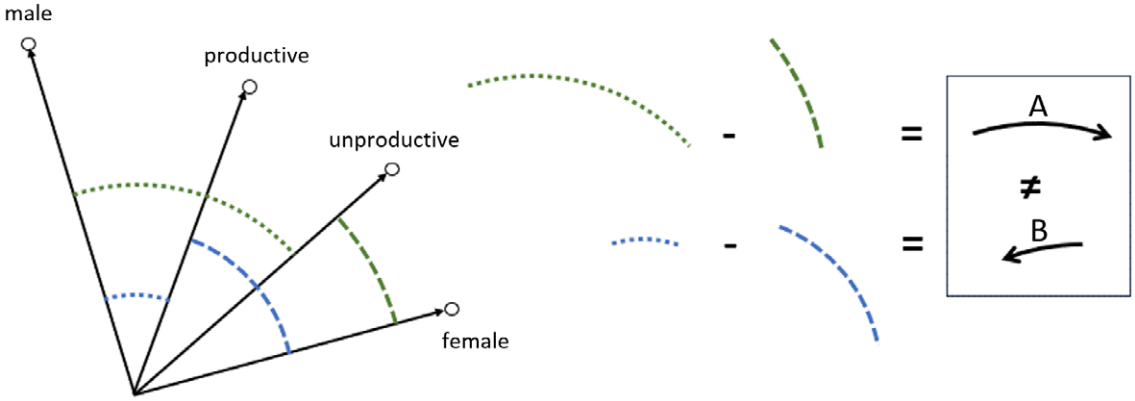
An unsolved issue in the domain of NLP is the perpetuation of stereotypical biases inherent in the training data. This has led to increased attention in the research community, but the focus has predominantly been on English models, often neglecting models for other languages. In order to better understand bias across languages, this thesis investigates bias in German, Dutch, and multilingual word representations. These vector representations capture semantic meaning in a numerical representation and are key components of many NLP applications.

Method

Preserved biases in word embeddings are represented by tighter vector associations between certain genders or races and their stereotypical characteristics or occupations. The Word Embedding Association Test (WEAT) is a metric that takes advantage of this fact by measuring the angle (cosine similarity) between different concepts (see figure below). For example, it measures how ‚productive' relates to male & female (green in the figure below, resulting in A) and compares this to how ‚unproductive' relates to male & female (blue in the figure below, resulting in B). The greater the difference between these two results, i.e. between A and B, the more likely it is that the embedding is biased towards the concepts in question. This approach is used on two types of word representations: Static word embeddings, which capture the meaning of the word itself, and contextualized word embeddings, which take into account surrounding words. We use Fasttext as our pre-trained static word embedding because it is available for multiple languages, allowing us to test our German and Dutch word lists. For the contextual word embeddings, we use three different models. The Bidirectional Encoder Representations from Transformers (BERT), a widely used word representation in research, will be investigated in its German, Dutch, and multilingual versions. In addition, the state-of-the-art multilingual models from OpenAI (released January 2024) and VoyageAI (released June 2024) are considered.

Data & Results

The data used for the bias metrics is partly derived from workshops conducted in the context of the EU-Project BIAS. The workshops aim to identify biases in the labor market and were held in the Netherlands to identify Dutch-specific biases, and in Switzerland to uncover German-specific biases. They involved experts from different fields, including human resources, NGOs, and machine learning. The results show that the biases identified in the workshops are reflected in almost all multilingual models, most German models, and some Dutch models.



Principle of Word Embedding Association Test (WEAT), shown in a hypothetical 2-dimensional word embedding space.