Languages and Gender bias

Proposer(s) / Proposatzailea(k): Olatz Arbelaitz, Xabier Arregi, Olatz Arregi, Olatz Perez de Viñaspre

Contact / Kontaktua: olatz.perezdevinaspre@ehu.eus

Description / Deskribapena

Systems built using machine learning techniques replicate the patterns that appear in the data used to learn. Nowadays, machine learning is used in many natural language processing contexts. Texts, as the society generating them, often have gender bias; they contain some combination of cultural, educational, gender, race, or other biases. Consequently, systems built applying machine learning techniques to these texts will also be biased systems. For instance, Garg et al. show that word embeddings (vector representation of words) used in Natural Language Processing (NLP) convey 100 years of gender and ethnic stereotypes and demonstrate how dynamics of embeddings reflect the change of stereotipical attitudes towards women and ethnic minorities.

Currently the use of machine learning based systems to make decisions in many companies and aspects of our lives is increasing, and if we don't get this right, we could be making wrong decisions that have critical consequences to someone's life. Consequently, these systems will contribute to increase the bias already existing in society.

This is something we should try to eliminate or even use in the opposite sense.

In this work, we will analyze the correlation between language and gender bias, by means of comparing different language models and their bias.

Goals / Helburuak

- Analyze gender bias in different languages (Language Models, monolingual and multilingual)

- Compare the results among languages

- Depending on the student interests and skills, we will focus on the analysis or on the identification and elimination of the bias

Tasks and plan / Atazak eta plana

- Study literature
- Analyze and decide the datasets and models to use
- Experimentation and evaluation
- (Optional) Write and submit a research article
- Write down the master thesis