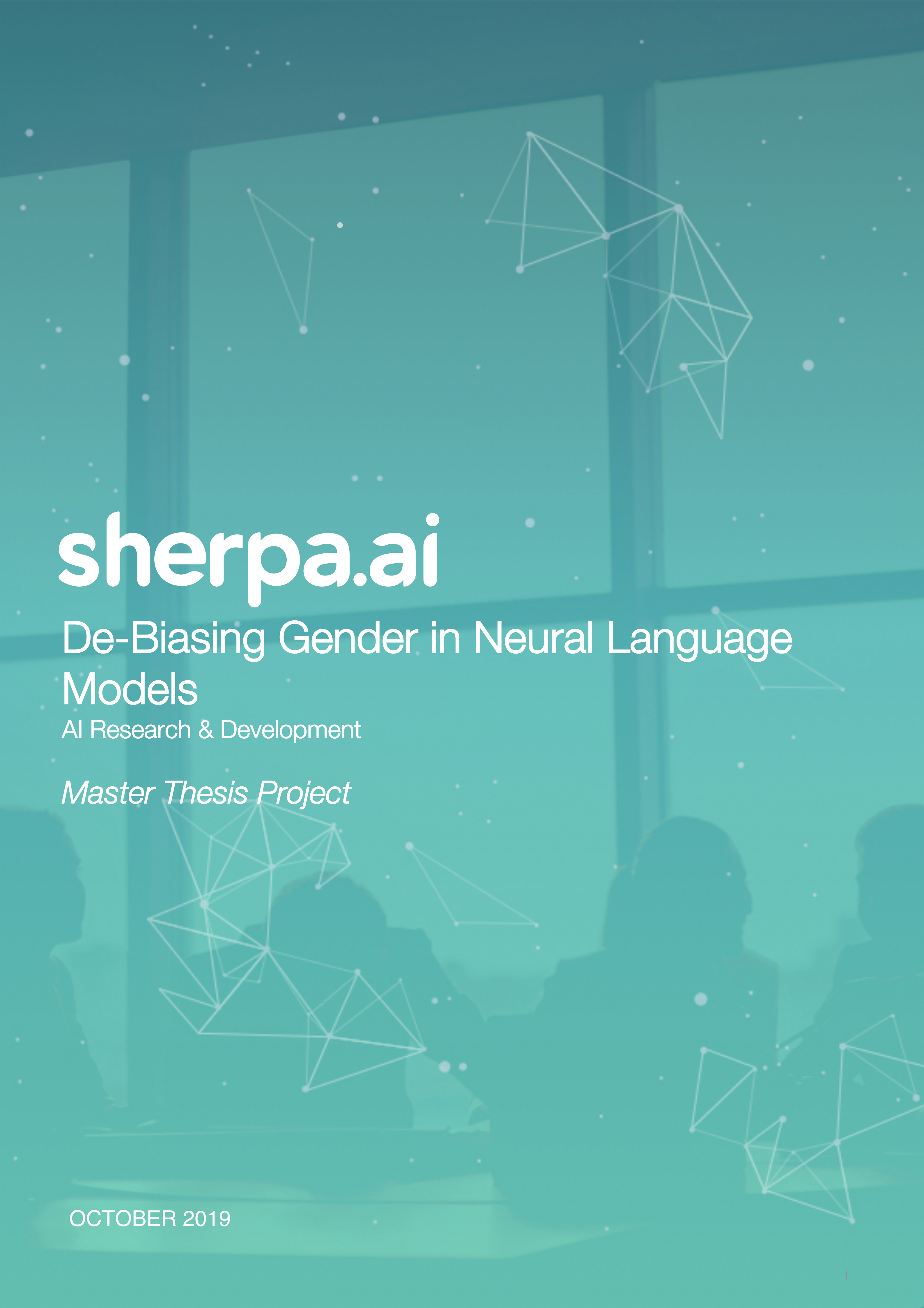

sherpa.ai

De-Biasing Gender in Neural Language Models

AI Research & Development

Master Thesis Project

OCTOBER 2019

Proposers and Contact Info (Academic and Industrial Mentoring)

- Dr. Miguel A. Veganzones (Sherpa AI Director): ma.veganzones@sher.pa

- Prof. Eneko Agirre (IXA Team, EHU/UPV): e.agirre@ehu.es

Project Description and Goals

Word embeddings and neural language models are widely used in NLP for a vast range of tasks. It was shown that word embeddings and neural language models derived from text corpora reflect gender biases in society. This phenomenon is pervasive and consistent, causing serious concern. **The aim of this project is to analyze the gender bias of the word embedding and neural language models employed in Sherpa and evaluate their effects in different tasks and languages.**

The gender bias of a word *w* has been defined by some authors as its projection on the “gender direction”:

$$w \cdot (he - she),$$

assuming all vectors are normalized. The larger a word’s projection is on *he – she*, the more biased it is. They also quantify the bias in word embeddings using this definition and show it aligns well with social stereotypes. A recent work claims that the bias is much more profound and systematic, and that simply reducing the projection of words on a gender direction is insufficient: it merely hides the bias. **The project consists of exploring state-of-the-art word embeddings and neural language models in English and Spanish to quantify the gender bias before and after applying de-biasing techniques. The project will evaluate the impact of the gender bias in News recommendation systems.**

Working Plan & Expected Results

1. Exploratory analysis of some state-of-the-art word embeddings and neural language models of interest in English and Spanish

2. Analysis of the gender bias in said models before and after applying de-biasing techniques

3. Analysis of the impact of gender bias in News recommendation systems

4. Report

Requirements

Basic knowledge of:

- NLP

- R / Python

Interest on:

- Neural language models

- Gender biases

Benefits and Practical Information

- Funding: 650€ / Month

- Duration: 3 – 6 Months

- Location: Aula SHERPA, Fac. Informática San Sebastián

Sherpa Europe, S.L. (Sherpa) accepts no liability for the content of this document, or for the consequences of any actions taken on the basis of the information provided. This document is intended to provide preliminary guidance in anticipation of further discussion and has not been prepared with the level of due diligence and analysis that would be needed to constitute a commitment of Sherpa. Anyone who receives this document are cautioned to consider that its contents are unaudited and that it may contain inaccurate, incomplete or summarized information, which may be change without notice. This document is confidential and its contents may not be totally or partially disclosed or reproduced, without the prior written consent of Sherpa. By allowing you access to the aforementioned document we shall have no liability, duty or obligation of any kind to you.