Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Nazmul Hasan
Assignment title:	part6
Submission title:	PR paperwork
File name:	ds_Reducing_Gender_Bias_in_South_Asian_Language_Transl...
File size:	230.67K
Page count:	6
Word count:	4,180
Character count:	22,242
Submission date:	09-May-2023 12:32AM (UTC+0530)
Submission ID:	2087839812

Towards Reducing Gender Bias in South Asian Language Translations

Nazmul Hasan Wanjan
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
nazmul.hasan.wanjan@g.bracu.ac.bd

Basit Hussain
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
basit.hussain@g.bracu.ac.bd

Malika Muradi
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
malika.muradi@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat@bracu.ac.bd

MD. Humaion Kabir Mehedi
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

MD. Mustakin Alam
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

***Abstract*—Gender bias in neural machine translation is not a new topic. However, works on reducing such bias from Natural Language Processing tasks are fairly new. In this paper, we have tried to analyze the gender bias that is observed during Machine Translation, one of the most notable applications of Natural Language Processing, of three of the South Asian languages and have tried to reduce that by one of the pre-existing methods, specifically by using more data for the biased segment of the population. The purpose of this paper therefore is to help the current machine translation systems reach a better state with more accurate prediction of the subject's gender with lesser dependency on other factors (mostly occupation) that are often unrelated.**

***Index Terms*—Gender Bias, Machine Translation, Natural Language Processing, Bias**

I. Introduction

Machine Learning fairness is a relatively new area which studies how to reduce favoring any specific segment of the population in machine learning models that may be the consequence of data or model inaccuracies. The end goal is to ensure equality for every segment of the population when they interact with the systems developed using machine learning. Older systems suffer from the problems of wrongly identifying segments of the population either due to the algorithmic or the dataset biases that are presented to it during the time of training. These biases, either coming willingly or unwillingly, hamper the modeling process and make it less generalized and usable for the whole population it is designed for. Additionally, it runs the risk of creating conflict among individuals in the real world as these models are often designed with the goal of being used in the real world systems.

Of the topics that concern machine learning fairness, bias reduction is a notable one. It is bias that leads us to stereotyping demographic groups, suppressing some and even leading the individuals within those demographic groups to believe certain aspects about themselves which are often misleading and wrong. Bias is an umbrella term to define the pre existing beliefs or shortcuts that may give a certain segment of the population some sort of edge over others. This has many types, one being gender bias (which is closely related to the topic of discussion of this paper), that comes from the beliefs about certain genders that exist in the society. For instance, historically, the majority of the existing cultures have been patriarchal. This means that the chief earning member is the father and he is also the head decision maker. Due to that, women are viewed as the segment who are weak and not suited to do physically and mentally challenging jobs. These pre-existing and dated thoughts, although useful in some situations in the past, are getting more and more inaccurate in the modern world where women, as well as other segments of the population, are joining the workforce for these demanding jobs.

In this work, we have tried to peer into the gender bias that may arise from the faulty and incomplete data and algorithms used for training the machine translation models, which is one of the most popular areas of Natural Language Processing. We specifically look into three of the South Asian languages: Bengali, Dari, and Urdu, and measure the gender bias that may occur when translating to and from English. Then we try