Ensayo sobre el debate de IA

Alejandro Santorum Varela
alejandro.santorum@estudiante.uam.es

14 de abril de 2019

Introducción

Llegado el final del debate de Inteligencia Artificial podemos resumir, analizar y comentar brevemente las participaciones en el foro de los estudiantes. El objetivo de este ensayo es hacer una breve reflexión sobre los sesgos de género que existen en los diferentes sistemas y/o algoritmos de inteligencia artificial.

Prejuicios de género en los sistemas de Inteligencia Artificial

Muchos de los debates en el campo de inteligencia artificial tratan sobre los prejuicios que parecen mostrar los dispositivos que se favorecen de algoritmos de aprendizaje automático. Son bien conocidos los ejemplos de los asistentes personales Siri, Alexa o Cortana, todos con voz de mujer que parece indicar que los humanos tendemos a pensar que una vez femenina es más servicial que la femenina.

Es importante que los ingenieros informáticos se fijen en estos debates para que estas consecuencias negativas no continuen repitiéndose. En un primer momento los ingenieros no pensaron que esto podía pasar, que los algoritmos

de aprendizaje automático que estaban alimentando con cantidades ingentes de datos no acogerían los sesgos machistas y prejuicios que existen en la sociedad actual, muchas veces porque confiaban en la sobre-neutralidad del algoritmo y, otras muchas veces, en que no nos paramos a pensar si ciertos conjuntos de texto contienen realmente prejuicios y discriminaciones.

Algunos ejemplos de posibles sesgos de género en los textos (principal fuente de datos para los algoritmos de procesamiento natural del lenguaje) son los siguientes:
La frecuencia con que aparece la palabra 'Mr' en texto anglosajones es superior que las palabras 'Mrs, 'Miss' and 'Ms' juntas [1]. Adicionalmente, las menciones individuales masculines duplican a las menciones individuales femeninas en los diferentes artículos científicos, sociológicos, literarios, etc [2]. En base a esto, un sistema de cuotas para el equilibrio de género en los datos de entrenamiento para algoritmos de aprendizaje automático puede servir para combatir gran parte del sesgo latente en las fuentes de datos de entrenamiento basadas en texto.

Para concluir, identificar el sesgo de género en los datos de entrenamiento para los algoritmos de aprendizaje automático es una tarea compleja pero no insuperable. El hecho de que los algoritmos de aprendizaje automático puedan aprender el sesgo de género puede ser muy interesante para su investigación de cara a enteder su prevalencia en la sociedad y así intentar mejorarla, no obstante, esto no es una ventaja en aplicaciones prácticas que toman decisiones sobre la vida de las personas.

Bibliografía

[1] Suzanne Romaine et al. 1998. Communicating gender. Psychology Press.

[2] Michael Pearce. 2008. Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. Corpora 3, 1 (2008), 1–29.

[3] Sandra L Bem and Daryl J Bem. 1971. Training the woman to know her place: The social antecedents of women in the world of work. Department of Psychology, Stanford University.

[4] Janet Holmes. 2002. Gender identity in New Zealand English. Gender across languages. The linguistic representation of women and men 1 (2002).

[5] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In Advances in Neural Information Processing Systems.