Chloe Hutchins, Monica Kavthekar and Jennifer Wright
Senior Design Writing 2
Due: 10/1/2019

Proposal For Bias Detection Web Plugin

Machine learning seems to be the impartial answer to problems when it comes to dealing with bias and prejudice. The idea that a machine could make decisions seems like the perfect way to make completely fair decisions. Upon closer investigation, it becomes clear that machine learning algorithms are not as impartial as people had once thought. The nature of machine learning algorithms includes being trained by past data, which means they reflect the biases that are shown in the data. This is a well known issue, but a hard problem to solve because it is at the core of how machine learning works. In order to ensure that machine learning algorithms are not making biased decisions, developers cannot train impartial algorithms with data that reflects all past biases.

In order to fix the issue of bias in machine learning, we must use machine learning to detect bias in online text. Therefore, in the future, internet content will not have so many of these issues. The few existing solutions for this problem involve framework and algorithms for altering the data before using it to train algorithms. They also focus more on training data versus recognizing bias in data. There are not any current solutions that involve visual representation of bias on a web page. Frameworks like FairML aim to create less biased algorithms by weighing the importance of each feature the algorithm is trained on and attempting to remove any words that indicate bias. Other tools include Audit AI, Google's What If, and IBM's AI Fairness 360 Open Source Toolkit. Although all of these are interesting tools, they all focus on improving our algorithms using biased data. Our web plugin for bias detection takes a different angle: analyzing the sources themselves. Our product aims to be easy to use, so anyone who wants to analyze the amount of bias in a given text can use it. The two step process of clicking on the web plugin is designed to create a minimum amount of friction, so the user is more likely to try it. Our product also provides a visual aspect that frameworks and algorithms do not. Creating a visual aspect means that results from the web plugin for bias detection are easy to interpret. There really are no other tools like our web plugin for bias detection on the market right now.

TransitionToday, Americans are lucky to live in a world where people care about moving towards equality. The unlimited resources on the world wide web help guide this push. Human bias often slips into the text corpora on the internet. According to Aylin Caliskan in her report *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*: "text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names". Bias towards race and gender is problematic, and indicates an implicitly biased set of resources on the internet. When it is unclear where that bias exists, it will be extremely challenging to eliminate the bias. Our web plugin aims to solve this problem. With our product, we aim to bring awareness to the bias that surrounds us on the web every day in a visual and easily interpretable way. Our tool will use machine learning to extend the research that Caliskan has done by using the Implicit Association Test to replicate biases and detect them. The tool will act as a litmus test, highlighting potentially biased phrasing and giving a general summary of the amount of bias in the work. Our hope is that with more awareness of the bias in work we rely on, content consumers will be able to recognize problematic biases and content creators will be held accountable to create less biased work in the future.

TransitionOne of the reasons our team wants to develop this product is because it has been found that males are more associated with the subjects of math and science while females are more associated with humanities-based subjects. An interesting study performed at Stanford asks the question: "why focus on the gender gap in math where males outperform and outnumber females over the gender gap in other subjects where females outperform and outnumber males?". However, math, according to *Explaining the Gender Gap in Math Test Scores: The Role of Competition* by Niederle and Vesterlund, has been shown as a good predictor of future income. Young women are learning from our biased web, where they see males being primarily associated with math. Therefore, they will think that there is no one like them in math, and thus, it is not a place for them. Before young women even know if they enjoy math, they are turned away from it. This implicit bias indicates a butterfly effect of bias leading to a pay gap. Being aware of the issue and how broad it is can be the first step in fixing it. Our web plugin for bias detection accomplishes this goal.

Our web plugin can be used by many different groups because bias is an issue that is ingrained into our society and has a wide-reaching influence. Our web plugin will allow companies to evaluate the bias in the information that exists regarding them online. The plugin will allow students to evaluate the bias in the information they use to research for school. Finally, everyday adults, many of whom get their news online, can use our plugin to evaluate the bias in the information that they are consuming on a daily basis.

This product is meant to be accessible. To make it so, the plugin will be available for download free of charge. One way this project can generate revenue is through the addition of extra features which clients can purchase. These extra features will be targeted at companies who may want to use this technology in order to evaluate and improve the way they are perceived online. These extra features will contain evaluation and statistics. However, the basic, invaluable information that comes from the plugin is meant to be shared. Bias is something that everyone encounters, whether they are aware of it or not, and this product intends to make people informed about the biases that they are consuming. Therefore, they will be able to better evaluate the information that they access online. Another way in which this product will be accessible is that it will be easy to use. This is because the product is targeted towards a wide variety of users who are not all likely to be technologically savvy. This product will just require users to download the web plugin, and then our algorithm will do all of the work. The visualizations will be organized in a way that can be easily understood by users. The product will highlight words and sentences that may be biased, as well as provide graphs and statistics describing the biases on the page. It is important to make this project accessible because it will have a greater impact. The purpose of this product is to confront the issue of bias in our society by making people more aware. Our product will create awareness because it will give the users statistics and visuals which show the bias that is contained in the web pages they are accessing. There is so much information online, and spreading our web plugin to more users will spread awareness to a greater population. How do you intend to create awareness among your target audience?

The broader societal need that a bias-detection web plugin will fulfill is the journey towards the elimination of discrimination online. Discriminatory, bias-ridden language comes from long-held human stereotypes, and its perpetuation on the internet leads to its continuance in society. Specifically, our plugin will concentrate on identifying bias in male vs. female language in text on the internet. Because these stereotypes are commonly known, we plan to use the implicit association test in our algorithm to identify the distance between pronouns and the gendered stereotypes they are often associated with. Then, we will create a visualization tool that will highlight portions with heavy bias on a webpage, and display graphics

about the total bias contact in a certain piece of text on a webpage. Users will be able to easily see the amount of biased content they are viewing, and they will be less influenced by the stereotypes. We will use bias visualizations to address the broad societal need of heavy stereotyping and bias in society.

Stereotyping is clearly a broad societal problem that exists in many forms and must be addressed in today's society. If our plugin was used widely, it would help decrease the many ways that stereotypes have affected human's daily lives, especially through the increased presence of the internet in today's society. Because of the omnipresence of the internet in human lives, the intrinsic bias that has been contained in human society since the beginning of time is now everywhere. Bias can be spread at a faster pace than ever before and can reach millions of people. According to an article published by the Association for Computing Machinery, a remedy for bias must start with awareness that bias exists. While creating awareness does not eliminate bias itself, it will guide society towards a solution. Programs like affirmative action are other examples of how bias visualization can allow for positive steps towards decreasing discrimination in broader society. In the realm of college education, scholars cite that low-income, racially underrepresented students have directly benefited from affirmative action programs. The programs have also worked to create a more diverse student body in all elite institutions of higher education. Diverse students bring their perspectives to the table, which will continue to battle stereotypes and work towards eliminating them altogether. Similarly, our web plugin will work to raise awareness of bias in content on the internet.

Because the internet can reach everyone in the world, the wide use of the web plugin will mean that it will be used internationally. Currently, in the west, many stereotypes of a patriarchal society that American society eschews are still commonplace in other places. Often, stereotypes are ingrained in culture, and people may not yet be open to changing their mentalities. In addition, bias may be viewed through different lenses by people from different countries. When creating the plugin, we will use western ideals to train the machine learning algorithm. Inherently, our biases will be evident in the first iteration of the web-plugin, giving it a limited scope. However, as the plugin expands, calling on diverse perspectives to allow for the algorithm to tailor itself to varying mentalities will allow for it to have a more international scope. In its first iteration, the plugin may hold some western ideals that are shared among citizens of the US and other similar societies. However, in areas where culture is dominated by religion or tradition, these values may not be embraced in the same way. Regardless, the plugin can certainly benefit the international world. In order to ensure that different cultures are treated with sensitivity, it will be important to tailor the plugin to international audiences. The approaches to eliminating stereotypes may be different internationally, but the desired result is the same- allow for internet content to be viewed through an unbiased lens. Does your project need regulation? Can it be misused?

Resources

Semantics derived automatically from language corpora contain human-like biases by Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan

Explaining the Gender Gap in Math Test Scores: The Role of Competition by Muriel Niederle and Lise Vesterlund

https://www.technomancers.co.uk/2018/10/13/five-tools-for-detecting-algorithmic-bias-in-ai/

https://cacm.acm.org/magazines/2018/6/228035-bias-on-the-web/abstract

https://www.gse.harvard.edu/news/uk/18/07/case-affirmative-action