[MUSIC PLAYING]

AUDACE NAKESHIMANA: In our work on fairness and AI, we present a case study on natural language processing titled "Identifying and Mitigating Unintended Demographic Bias in Machine Learning." We will break down what each part of the title means. This is the work that was done jointly by Chris Sweeney and Maryam Najafian.

My name is Audace Nakeshimana. I am a Researcher at MIT, and I'll be presenting their work. The content of the slides presents a high-level overview of a thesis project that was done throughout a course of the year. It will be released soon on MIT DSpace.

AI has the power to impact society in a vast amount of ways. For example, in the banking industry, many companies are trying to use machine learning to figure out if someone will default on a loan given the data about them. Now, because machine learning is used in the high-stakes applications, errors that cause it to be unfair could cause discrimination, preventing certain demographic groups from gaining access to fair loans.

This problem is especially important to address in developing nations where there may not be existing sophisticated credit systems. Those nations will have to rely on machine learning models to make these high-stakes decisions, such as alternative credit scoring mechanisms that are possibly going to be involving AI more and more.

This work focuses on applications of machine learning in natural language processing. NLP is important to studying fairness in AI because it is used in many different domains, from education to marketing. Furthermore, there are many sources of unintended demographic bias in the standard natural language processing pipeline. Here we define the NLP pipeline as a combination of steps involved, from collecting natural language data to making decisions based on the NLP models trend and resulting data.

Lastly, [? therefore, ?] natural language processing systems is easier to get. Unlike

tabular systems from banking or health care, where companies may be reluctant to release data due to privacy concerns, NLP data, especially in widely spoken languages like French or English, is available from different sources, including social media and different forms of formal and informal publications, making it more effective to use in research on how to make NLP systems more fair.

We now break down what unintended demographic bias means. The unintended part means that this bias comes as an adverse side effect, not deliberately learned in a machine learning model. The demographic part means that the bias translates into some sort of inequality between demographic groups that could cause discrimination in a downstream machine learning model.

And finally, bias is an artifact of natural language processing pipeline that causes this unfairness. Bias is [INAUDIBLE] term. Therefore, it is important that we center on a specific form of bias that causes unfairness in typical machine learning applications. In gender-based demographic bias, for example, machine learning model might associate specific types of jobs to specific gender just because it's the way it is in the data used to train the model.

Within unintended demographic bias, there are two different types of bias that will focus on in natural language processing applications. These are bias in sentiment analysis systems that analyze positive or negative feelings associated with words or phrases and toxicity analysis systems designed to detect derogatory or offensive terms in words or phrases.

Sentiment bias refers to an artifact of the machine learning pipeline that causes unfairness in sentiment analysis systems. And toxicity bias is an artifact of the pipeline that causes unfairness in systems that tries to predict toxicity from text. In either sentiment analysis or toxicity prediction, it is important that our machine learning model doesn't use sensitive attributes describing someone's demographic to inform them whether a sentence should be positive or negative sentiment or toxic or less toxic.

Toxicity classification is used in a wide variety of applications. For example, it can be used to censor online comments that are too toxic or offensive. Unfortunately, these algorithms can be very unfair. For example, the decision of whether sentence is

toxic or non-toxic can depend solely on the demographic identity term, such as American or Mexican, that appears in the sentence.

This unfairness can be caused by many different artifacts of the natural language processing pipeline. For instance, certain nationalities and ethnic groups are specifically more frequently marginalized. And this is reflected in the language usually associated with them. Therefore, training NLP algorithms and resulting data sets could result in a certain form of unintended demographic bias.

We want to drive home the point of unintended demographic bias versus unfairness. Unintended demographic bias can enter a typical machine learning pipeline from a wide variety of sources, from the word corpus to the word embedding, the data sent to the algorithm, and finally from the thresholds used to make decisions. The possible unfairness or the discrimination comes at the point where this machine learning model meets society and actually causes harm. This work addresses mitigating and identifying unintended demographic bias at each stage in the natural language processing pipeline, from the words corpus to the decision level.

Our big goal here is to find ways to mitigate the bias that we might inherently find in the text corpora or other types of data representation that are used to build NLP applications. For this module, we cover measuring unintended demographic bias in word embeddings and using adversarial learning to mitigate word embedding bias. The corresponding thesis goes further, and it covers techniques for identifying and mitigating unintended demographic bias at other stages of the NLP pipeline.

We now cover the work as measuring word embedding bias. Word embeddings [? encode ?] [? text ?] into vector spaces where distances between words describe a certain semantic meaning. This allows one to complete the analogy of man is to woman as king is to queen.

Unfortunately, researchers Tolga Balukbasi and others found that even for word embeddings trained from Google News articles, there exists bias in word embedding space, where the analogy becomes man is to woman as computer programmer is to homemaker, another word for a housewife. This is concerning given that word embeddings could be used in natural language processing applications devoted to

predicting whether someone should get a certain job. However, it is difficult to quantify the bias just based on the vector space analogies.

In this work, researchers Sweeney and Najafian develop a system to measure sentiment bias in word embeddings to a specific number. The way they do this is they take the bias toward embeddings and use them to initialize an unbiased labeled word sentiments data set. They train a logistic regression classifier on this data set, and they predict negative sentiment for a set of identity terms.

For example, in this case, this is a set of identity terms describing demographics from different national origins. They analyzed the negative sentiment for each identity term and predict a score that describes the bias in word embeddings. This score is the divergence between the [INAUDIBLE] for abilities, for negative sentiment, for national origin, identity terms, and the uniform distribution. The uniform distribution describes a perfectly fair case, wherein a demographic is receiving an equal amount of sentiment in the word embedding model.

Now that we have a grasp on the word embedding bias, we can start to figure out how to mitigate some of this bias. In the thesis, Sweeney and Nafajian describe how they use adversarial learning to debias word embeddings. Different identity terms can be more or less correlated with positive or negative sentiment. For example, words like American, Mexican, and German can have more correlations with negative sentiment subspaces and positive sentiment subspaces, because in the data sets used, it might appear to be more frequently associated with negative or positive sentiments.

This is concerning. Even downstream machine learning model picks up on these correlations. Ideally, you want to have each of those identity terms to a neutral point between negative and positive sentiment subspaces without distorting their meaning within the vector space so that the word embedding model can still be useful. They use an adversarial learning algorithm to achieve this. More details of this algorithm are described in the corresponding phases.

I now present some of their work in evaluating how adversarial learning algorithms can debias word embeddings and make the resulting natural language processing system more fair. We focus on realistic systems in both sentiment analysis and

toxicity prediction. For each application, Sweeney and Najafian define fairness metrics to let us know whether the debiased word embeddings are actually helping.

These fairness metrics often come in the form of a templates data set. Researchers have created these data sets to somewhat tease out different biases with respect to different demographic groups. For example, this set is meant to tease out biases between African-American names and European-American names when substituting each name out in the same sentence. Similar template data sets have been created for toxicity classification algorithms, where you sub out different demographic identity terms within a sentence and compare differences in the overall toxicity predictions.

Sweeney and Najafian used these templates data sets to compute fairness for a real-world toxicity classifier. This graph shows per-term AUC distributions for CNN convolution neural network that was trained on a toxicity classification data set. The x-axis represents each demographic group, where the templates data set has that identify term subbed in for each sentence. Each dot describes a particular training run of the CNN.

The y-axis describes the area under the curve accuracy for this template's data set. One can see that there is a lot of disparity between the accuracies for different demographic groups. Ideally, you would want the variance in different training runs to be compressed as well as the differences between each demographic group in the AUC scores to be smaller. Sweeney and Najafian show a toxicity classification algorithm that uses the debias towards embeddings creates better results.

This slide shows results for per-term AUC distributions for the CNN with different debias treatments. Sweeney and Najafian measure how their word embedding debiasing compares to other state-of-the-art techniques. Further discussion and evaluation of these graphs are presented in the corresponding thesis.

To wrap up, we describe some key takeaways from this project. First, there is no silver bullet. There are many different types of applications and various types of bias to correct for when trying to make NLP systems more fair. Second, bias can emanate from any stage of the machine learning pipeline. Therefore, having to also identify and mitigate bias at all stages of the machine learning pipeline is essential.

Finally, we focus on solving this problem within an academic context for natural language processing pipeline, but this cannot all be solved in academia. For example, much of the unintended bias in the data set, like the text corpus, could come from decisions made upstream in direct collection. Furthermore, unintended bias could come from decisions made when deploying the model into society.

When the model is used in a way that does not resonate with how the data was collected in the first place, this could cause discrimination. An example of this is when the data collected from a specific demographic population is used to make predictions that affect other demographics that were not taken into account during data collection. Finally, it is important to have efficient channels of feedback for these machine learning models.

The work presented in this module highlights why fairness is a very important concept. It is therefore critical for data scientists and engineers to measure and understand performance of their models not just through accuracy, but also through fairness.

[MUSIC PLAYING]