School of Informatics

Informatics Project Proposal
How the Geometry of Embedding Spaces Affects Bias in NLP Systems

B154805
April 2020

Abstract

Determine the validity of bias metrics and debiasing techniques that rely on modifications to word embedding spaces, by correlating changes in word embedding space to bias in standard downstream NLP applications. EDIT THIS GREATLY

Date: Wednesday 22$^{\text{nd}}$ April, 2020

Tutor: Resul Tugay
Supervisor: Seraphina Goldfarb-Tarrant

1 Motivation

With the recent proliferation of the use of machine learning for a wide variety of tasks, researchers have identified unfairness in ML models as one of the growing concerns in the field. Many ML models are built from human-generated data, and human biases can easily result in a skewed distribution in the training data. We build intelligent systems that learn enough about the properties of language to be able to understand and produce it, in the process it will also acquire historical cultural associations, some of which can be objectionable. Already, popular online translation systems incorporate some of the biases we study (see the supplementary materials). Further concerns may arise as AI is given agency in our society. If machine learning technologies used for, say, résumé screening were to imbibe cultural stereotypes, it may result in prejudiced outcomes. ML practitioners have recently become proactive in recognizing and counteracting these biases, to prevent our models and products from perpetuating unfairness by performing better for some users than for others. Bias in Natural Language Processing systems has recently received a great deal of attention in the research community. The majority of approaches to measuring bias do so via examining the spatial relationships between word embeddings. For example, some characterize Gender as a direction in the word embedding vector space. These spatial relationships have been shown to reflect and to be predictive of many human biases.

There is considerable research on mitigating bias in NLP systems by altering word embeddings using various techniques. While such debiasing has been shown to be effective for some target criteria that measures bias, Gonen et al (Lipstick on a pig) say that the actual effect is mostly hiding the bias, not removing it. Bias information is still reflected in the clustering pattern of the "gender-neutralized" words in the debiased embeddings, and can be recovered from them. They observed that most word pairs maintain their previous similarity, despite their change in relation to the gender direction. The implication of this is that most words that had a specific bias before are still grouped together, and apart from changes with respect to specific gendered words, the word embeddings' spatial geometry stays largely the same.

It is currently unknown how much the differences in spatial relationships really affect practical downstream NLP applications. There is existing work which shows that fine-tuning word embedding spaces can affect downstream tasks. However, there is no systematic study of how the different notions of bias impact diverse downstream tasks.

1.1 Problem Statement

It has been shown that word embeddings contain human stereotypes and bias. To counter this researchers have been working on modifying embeddings to get rid of the bias. Bias can be seen as spatial differences between clusters in the vector space. These clusters represent groups of words having some attribute in common (eg: cluster of words deemed feminine - pretty, elegant, delicate, soft, cute). Debiasing techniques at the moment modify the geometry of the vector space such that the magnitude of the distance between clusters reduces. In some sense, the distance between clusters represents the intensity of bias. What we are trying to do is analyse how this reduction in intensity of bias affects downstream tasks as compared to original word embeddings. Our research tries to answer the following questions:

1. Do all spatial differences matter?

2. Does the magnitude of the distance between clusters of words relating to the bias attribute have a noticeable effect?

3. Does this differ by type of downstream task? Eg: toxicology, co-reference resolution, etc.

4. Does this differ by the type of embedding algorithm?

This project will involve training different word embeddings and implementing a standard NLP classification task. It will involve examining correlation between bias as measured geometrically in the original embedding, and as measured in the downstream task via standard metrics for classification bias.

1.2 Research Hypothesis and Objectives

We are going to study how much debiasing word embeddings can affect a downstream task. We will train GloVe (ADD REFERENCE) and FastText (ADD REFERENCE) embedding algorithms from scratch.

Our downstream task of choice is hate speech detection. We will evaluate the performance based on intrinsic and extrinsic metrics. For intrinsic metrics, we will rely on WEAT (ADD REFERENCE FOR Caliskan et al., 2017) For extrinsic metrics, we may use Equality of Opportunity (Hardt et al., 2016) or pinned AUC (area under curve) (Dixon et al., 2018)

(MOVE DETAILS OF WORD EMBEDDING ALGORITHMS AND EVALUATION METRICS FOR BACKGROUND TO HERE)

The study will not involve new debiasing techniques or comparison of language specific bias reduction effects. It is mainly an evaluation study.

1.3 Timeliness and Novelty

Of late, the field of fairness in machine learning has become very important. It is high time such a study is conducted because people have been using new techniques for debiasing systems but a thorough investigation of its effects has not been conducted. Our research will be the first study into the evaluation aspect of debiasing embeddings.

It is of great importance right now. Ethics and fairness in AI is blowing up and it is being used everywhere. It is necessary for us to see that these systems do not propagate and enhance the bias (ADD REFERENCE TO SOME PAPER). If such studies are not conducted, we could potentially be releasing new systems into the real world that seem to solve fairness issues superficially but are still inherently biased. – One such approach is seen in the nascent field of fairness in machine learning,

1.4 Significance

With NLP systems being used in all fields it is of utmost importance that these systems are fair. They need to be modeled in a way that allows equal opportunity to everyone. They shouldn't be biased towards or against any set of people. This need is being sated by research into ethics and bias. However, the new techniques have not been studied in terms of their effect on downstream tasks. It is important to do this evaluation before deploying systems that use these techniques in case we have not in fact got rid of everything we claim to have.

We believe that our study will help further research in this area as it give us some understanding of whether this approach is good enough and research should continue in this direction or there is need to develop new techniques that do more than hide the bias.

1.5 Feasibility

Our project pipeline is given below.

1. Collecting data: Existing datasets need to be identified as suitable to train the embedding algorithms.
2. Embedding algorithms: code base for GloVe and FastText already exist. We will need to fine-tune the code to fit our needs.
3. Downstream task: existing models for hate speech detection will be used.
4. Evaluation: Use WEAT and Equality of Opportunity to measure changes in bias in the downstream task

Thus, our baseline system shouldn't be hard to implement pretty quickly. We can run tests on this to get measurements to compare with.

Debiasing the embeddings will be our most time consuming task. However, except for this added task the pipeline remains the same. Therefore, we feel this project is feasible given the circumstances, time frame and resources.

1.6 Beneficiaries

No matter what results we get, we believe it will be relevant and benefit the NLP community. If we manage to prove that post-hoc editing of word embedding space is enough, more researches can employ the existing debiasing methods to downstream applications. If we prove otherwise, researchers will need to find ways to actually remove the bias not just hide it. Either way the community will know how to move forward. If we get clear results we will of course try to publish and present our work.

2 Programme and Methodology

Work Package 1

First we will need to set up and implement a baseline system. The results we get on running this system will be compared with the results of our debiased word embeddings system. We follow the pipeline mentioned in section 1.

1. **Collecting data:** The common corpora for training word embeddings are usually from Wikipedia, online magazines, and news articles. This is because we want a lot of context for words so that the word embedding algorithm can accurately model co-occurrences and

semantic relationships between words. (NOT SURE ABOUT THE PREVIOUS STATEMENT. DISCUSS WITH GROUP) Off-the-shelf models for hate speech detection do exist. However, not all of them make use of word embeddings. If we need to build our own classifier we will need data for this too. The corpus for this task would contain text from social media such as Facebook, Twitter, and Reddit. These platforms are usually what people use to express their opinions and hence they are perfect for our task.

This task will take us upto 1 week.

2. **Embedding algorithms:** We are going to use GloVe and FastText word embeddings for our project. We do this because we want to know if different embedding strategies model bias differently. Code bases for both already exist. FastText code is written in python and we will need to fine-tune the code to fit our needs. GloVe code is written in C++ but there should be python wrappers available. Again, we will need to fine-tune the code to fit our needs.

   Finding a suitable python wrapper might take more time than we anticipate so our worst case duration for this task is 2 weeks.

3. **Downstream task:** Hate speech detection is a fairly popular application in NLP so there are existing models for hate speech detection. We will make use of one of these for our work. However, if the model we select does not have a provision for word embeddings we will need to add this and retrain the model.

   Training and fine-tuning the model to the best settings will take up to 1 week.

4. **Evaluation:** We will use WEAT (implicit bias metric) and Equality of Opportunity (explicit bias metric) to measure the bias present in the hate speech detection classifier.(Are details needed in this section?)

   Familiarising ourselves with the two metrics and taking the actual measurements will take us 1 week.

Since some of are tasks are independent of each other we can perform them simultaneously. The time we have allotted to each task in this work package is highly exaggerated. However, if we do in fact take 5 weeks to do this we will still have enough time to complete the rest of the project.

Work Package 2

Next we need to set up our main experiment. This involves performing classification using debiased embeddings. The pipeline for this is given below.

1. **Debias embeddings:** We need to debias the word embeddings using one of the current debiasing techniques. We have decided to use dataset balancing. This involves either subsampling the dataset so that we have an equal number of positive and negative samples, or it involves generating synthetic data so that there are equal positive and negative samples.

   This task will take us 2 weeks because we will need to implement it from scratch. DISCUSS WITH GROUP

2. **Retrain embeddings:** Since we have already got suitable training algorithms for the word embeddings in our baseline system, we only need to feed it our new dataset. No other changes will be required.

   This task will take us <1 week.

3. **Downstream task:** We will use the same model we built in the baseline system but provide it with the debiased embeddings instead.

   This task will take us <1 week.

4. **Evaluation:** This task remains the same as the evaluation task in the baseline system.

   Since we will have understood the metrics by now, this task will take <1 week.

None of are tasks here are independent of each other so a delay in any pushes the timeline. However, the time we have allotted to each task in this work package is highly exaggerated. If we do in fact take 2-2.5 weeks to complete this we will still have enough time to complete the rest of the project.

Work Package 3

REWRITE THIS SECTION. BADLY WRITTEN
At this stage we will have all the results we need. We now need to compare them. We need to do the following.

1. Measure change in bias in the downstream task using GloVe
2. Measure change in bias in the downstream task using FastText
3. Compare how much the bias changes in our downstream task with another downstream task
4. Compare how much the bias changes in our downstream task when we use a different language.

THINGS TO DISCUSS WITH THE GROUP:

1. Should I mention point 3 and 4 in WP 3? They involve R&R's results. Am I allowed to discuss other people's work in my thesis?
2. Is the timeline okay?
3. Should I add a work package for actually writing my thesis or is it understood that the remaining time would be used for this?

2.1 Risk Assessment

We can categorise our risks into 3 types.

The first type of risk is that of not getting any conclusive results. This would push the understanding of whether current methods are a good way to debias NLP systems or not. If it isn't it could delay the development of better techniques. It could also lead to not superficially fixed systems to be deployed into the real world.

A more project-specific risk is that off-the-shelf models for hate speech detection may not have a provision for word embeddings. This could lead to a delay in our timeline as we would need to train our own model to be able to use word embeddings. Another risk is the use of FastText. Though it is a very efficient algorithm, it may not behave as we want it to in terms of either capturing language characteristics or debiasing techniques may lead to trouble in the downstream task model.

The final risk is that of failure to access resources. As we are now working remotely, it is possible that we may lose access to the university servers and GPU cluster.

2.2 Ethics

Since we are making use of existing and widely used corpora for our work, there is no ethical issue related to people's privacy. The topic of our project indeed contributes to ethical and fair use of machine learning techniques in society.

3 Evaluation

We require data that are natural language texts from sources such as Wikipedia, Reddit, articles etc. Well known datasets exist already for Wikipedia and Reddit texts. We will use one of these to train our embedding algorithms. size of corpus (which must be large enough to train embeddings, so minimum 10 7 tokens, or larger if you plan to subsample it). We can select an interesting corpus that we suspect might display some baseline level of bias - but almost everything will to some extent, so it doesn't really matter.

We plan to use off the shelf models for hate speech detection. However, in case we need to train our own model, we will need data from sources like Twitter, Reddit etc.There are common datasets that are used for hate speech detection purposes and we would most likely use one of those.

The results of our study are the changes in amount of bias in the downstream task after debiasing the word embeddings. We measure bis in the downstream task with implicit WEAT and explicit Equality of Opportunity metrics. (EXPLAIN BOTH AND REFERENCE PAPERS)

Once we have metrics, we need to modify the embedding spaces and observe how changes in WEAT effect changes in downstream metrics. Equality of Odds, proposed in [9], is a definition of fairness that is satisfied when the false positive rates and false negative rates are equal across comments containing different identity terms. This concept inspires the error rate equality difference metrics, which use the variation in these error rates between terms to measure the extent of unintended bias in the model,

4 Expected Outcomes

Since it is the first study of its kind we cannot predict what will happen. We are hoping to prove or disprove that hiding the bias is enough. It is an unexplored topic thus it will make an original contribution to knowledge.

Our initial instinct is that the bias in our downstream task will decrease but not enough to make the model more fair. However, we cannot say this is more likely than the other choice. DO THIS: Steal some part from where we spoke about filling gaps. it will help to better understand word embedding space, debiasing algorithms, how bias affects downstream tasks.

References