Gender Bias in Word Embeddings

Project Report

Ankit Pant – 2018201035
Tarun Mohandas – 2018201008
Supervised by: Dr. Manish Shrivastava

Abstract

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation. The blind application of machine learning on Web embeddings runs the risk of amplifying biases present in data. Word embeddings trained on a standard Natural Processing dataset, Google News Articles, exhibit female/male gender stereotypes to a disturbing extent.This raises concerns because their widespread use, often tends to amplify these biases.

The project aims to explore gender bias present in word embeddings and the techniques used to reduce/eliminate this. An attempt to explore other kinds of bias as well as other de-biasing would also be made.

Contents

1 Introduction

2 Literature Review

3 Methodology

4 Experimentation

5 Results

6 Conclusion

7 Future Scope

8 Introduction to Word Embeddings

Word embedding is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc. They are vector representations of a particular word. *Word2Vec* is one of the most popular technique to learn word embeddings using shallow neural network.

Word2Vec can utilize either of two model architectures to produce a distributed representation of words: continuous bag-of-words (CBOW) or continuous skip-gram. In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words. The order of context words does not influence prediction. In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words.

9 Introduction to Bias

It is important to quantify and understand bias in languages as such biases can reinforce the psychological status of different groups [4]. Biases differ across people though commonalities can be detected. A number of online systems have been shown to exhibit various biases, such as racial discrimination and gender bias in the ads presented to users. Different demographic and geographic groups also use different dialects and word-choices in social media. An implication of this effect is that language used by minority group might not be able to be processed by natural language tools that are trained on "standard" data-sets.

9.1 Gender Bias

Gender bias in language has been studied over a number of decades in a variety of contexts. Common biases link female terms with liberal arts and family and male terms with science and careers [3]. While there are more words referring to males, there are many more words that discriminates females than males. The following are some of common gender biases found in the standard datasets:

1. Man is to computer programmer as woman is to homemaker

2. Man is to doctor as woman is to nurse

Although the aforementioned examples prominently show gender bias, the following associations does not show any malignant bias and need not be modified.

1. Man is to king as woman is to queen

2. Man is to actor as woman is to actress

Since the aforementioned examples have gender specific modifications to the words, they do not show unhealthy bias towards any gender.

9.2 Other types of Biases

Some other kinds of biases[2] that are encountered in Machine Learning include:

9.2.1 Historical Bias

According to *Suresh et. al. (2009)*, "Historical bias is a fundamental, structural issue with the first step of data generation process and can exist even given perfect sampling and feature selection."

9.2.2 Statistical Bias

Difference between statistic's expected value and true value.

9.2.3 Unjust Bias

Disproportionate preference for or prejudice against a group.

9.2.4 Evaluation Bias

Biases in Benchmark Datasets (which spur on research).

10 Techniques for De-biasing

Generally the first step towards de-biasing the embeddings is to identify a subspace *called gender subspace* of the embedding that captures the bias[1]. After identifying the subspace, some of the techniques used for de-biasing include:

- **Neutralize and Equalize:** This is also called Hard de-biasing. *Neutralize* ensures that no gender-neutral words are present in the gender subspace and *Equalize* equalises the set of words outside the subspace [1].

- **Soft Bias Correction:** It is a linear transformation that seeks to preserve the pairwise inner products between all the word vectors while minimizing the projection of the gender neutral words onto the gender subspace [1].

11 Goals of the project

This project majorly aims at achieving the following goals:

- **Explore gender bias in word embeddings**: Explore the notion of gender bias in language(NLP datasets) and how it is captured in (associated with) word embeddings representation.
- **Recognise the gender bias in the dataset**: Given a corpus/word embedding, all unhealthy gender biasing needs to be recognized.
- **De-biasing the dataset**: After gender biases have been identified, they need to be de-biased, so that there is minimal bias when using the word embedding.

Once the aforementioned objectives are met, attempts would be made to do exploratory study to find other types of unhealthy biases existing in different word embeddings like *Word2Vec* and *ELMO*. Once this is completed, some creative and constructive solutions to de-bias other unhealthy biases found in different word embeddings like *Word2Vec* and *ELMO* would be explored.

12 Timeline

Week 1	Explore different word embeddings (Word2Vec, ELMO)
	Explore different pre-trained models and architecture used for training
Week 2	Explore Similarity and Analogy problem in NLP
	Input various types of biases and check if it shows up in word embedding
Week 3+	Analysis to check correlation between bias and data
	Trying different de-biasing methods on the different word embeddings
Bonus	Exploring other biases present in word embeddings
	Exploring solutions to de-bias the above biases

References

[1] Tolga Bolukabasi, et. al., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain

[2] Rachel Thomas, *Algorithmic Bias*, https://www.youtube.com/watch?v=pThqge9QDn8&list=PLtmWHNX-gukKocXQOkQjuVxglSDYWsSh9&index=16

[3] B. A. Nosek, M. Banaji, and A. G. Greenwald. *Harvesting implicit group attitudes and beliefs from a demonstration web site*. Group Dynamics: Theory, Research, and Practice, 6(1):101, 2002.

[4] E. Sapir. *Selected writings of Edward Sapir in language, culture and personality*, volume 342. Univ of California Press, 1985.