School of Informatics

Informatics Project Proposal
How Modifying the Geometry of Spanish Word Embeddings Affects Bias in Downstream Tasks

s1983961
April 2020

Abstract

Word embeddings are at the heart of the current research on natural language processing. However, they reflect and increment some of the biases that humans have. While some research has been made on how to measure and alter the intrinsic bias of these embeddings, none has been made on how that alters the bias on downstream tasks. It is also important to note that little to no research has been done on romance languages, particularly in Spanish. The aim of our work is to study and measure the relationship between the intrinsic and the extrinsic bias in Spanish word embeddings.

Date: Monday 27th April, 2020

Tutor: Amna Shahab
Supervisor: Seraphina Goldfarb-Tarrant and Adam Lopez

1 Motivation and Background

In recent times, Artificial Intelligence systems have been deployed in real-world applications at an ever-increasing speed. Neural networks and big data have been at the core of this rapid expansion. When talking about natural language processing, these systems deal with tasks ranging from translation to recognizing emotions and categorizing texts. These have been used on their own, such as Google Translate, or as part of larger systems, such as providing medical information to communities that speak low-resource languages [1].

However, some alarming concerns have been raised that we do not exactly know how these systems really function. One of these concerns is that machine learning systems tend to pick biases in the data used. This can be due to the actual content of the data [2] or due to biases that the annotators have [3].

An example for this are profiling systems that use data from previous arrests to inform whether someone is a likely suspect or not. These kinds of systems have been shown to be less accurate with people of color, sometimes showing alarming rates of false positives [4]. Another example of this are correferece systems, which tend to mirror gender stereotypes. In sentences mentioning a "he" and a "her" as well as a doctor and a nurse, these systems are more likely to state that the doctor is the "he" and the nurse is the "she", even if the rest of the sentence would indicate otherwise [5].

One of the places where we have found these biases reflected are word embeddings. These are vectors that represent words, which can then be fed into machine learning systems. Currently, the three most common embedding right now are word2vec [6], GloVe [7], and FastText [8]. While the embedding algorithm used doesn't affect much the result of the downstream tasks [9, 10], their training times and how they learn the words differ greatly. For our purposes, it is important to note that word2vec and GloVe use whole words to learn, while FastText uses subword information.

When Mikolov et. al. introduced the word2vec algorithm, they pointed out that the geometry of word embeddings tends to capture certain semantic and syntactic characteristics of the language. One of the examples they give is that if we take the vector for "king" and add the vector that goes from "man" to "woman", the result will be and embedding that is the nearest to "queen". However, Bolukbasi et. al. [11] noticed that when doing the same with "computer scientist", we would get "homemaker"!

Some attempts have been proposed to remove bias from word embedding algorithms, such as using PCA to identify and remove the dimensions which carry the most amount of data that could identify a given group [11]. Another proposal has been to balance datasets so that all relevant groups are represented equally [5, 2]. However, most of these attempts merely hide the biases instead of actually removing them, as bias tends to be highly interdependent with other variables and are sometimes deeply rooted into our cultures and languages [12]. This can be clearly seen when comparing word clusters: instead of removing them, these algorithms only bring them closer, instead of removing them altogether [13].

Another important thing to note is that, even though most of the recent studies have focused on the geometry space of the embedding spaces, as of yet no one has studied how altering these spaces effects downstream tasks. An important metric when doing this would be the level of equality of opportunity. That is, whether two groups of protected characteristics are treated equally fairly in a given task.

A relevant task in this regard is classifying whether a given comment is hate speech or not,

so that is the downstream task we will be focusing on. More specifically, we will be using the hatEval dataset [14]. This dataset has messages from Twitter in English and Spanish and is intended to be used to detect hate speech against women and migrants.

Finally, we know that languages with gendered pronouns, such as Italian, tend to form clusters around the gender of the of the words, rather than their semantic similarity [15, 16]. Given that most of the research done in this area has been relatively recent and that almost none of it has focused on romance languages, it would be interesting to study algorithms that reduce bias in a language such as Spanish, especially when comparing gender bias with other kinds of bias, such as race.

1.1 Problem Statement

We will be focusing on altering the bias found in Spanish word embeddings as measured intrinsically and the effect this has on an extrinsic measure on a downstream task. This task will be to classify whether internet comments are considered to be hate speech against women and/or migrants.

1.2 Research Hypothesis and Objectives

To measure the effects that reducing bias on the embedding space has on downstream tasks.

We expect both of our metrics to show reduced amounts of bias when compared to the unmodified embedding space.

1.3 Timeliness and Novelty

Even though the two papers that originated this field of study were published in 2016 [11] and 2017 [17], most of the papers have been published fairly recently, since 2018. Also, as mentioned before, most of the papers tend to focus on the embedding space itself and not so much on the downstream task.

As for the downstream task chosen, we are determining whether internet comments in Spanish are considered hate speech, either against women or against migrants. This is especially relevant for Latin-American countries, considering the waves of feminist strikes that have happened recently in the region [18, 19] and the recent migrant caravans that mobilized towards the United States [20, 21].

1.4 Feasibility

Setting up the general pipeline for the experiments should not be too time-consuming and we would expect to be able to run additional experiments, such as with different embedding spaces or different methods to remove the bias.

However, given the current state of the world and the uncertainty it carries around, we decided to err on the side of caution with our time estimates. Therefore, we are considering our minimum viable product to consist of the results of the experiments using the FastText embedding algorithm, along with two of the methods used to alter bias. This is explained in further detail in sections 2 and 5.

1.5 Beneficiaries

Biases tend to affect more adversely groups that have historically borne the brunt of discrimination, such as women, minorities and other groups at risk. The effects of these biases have been expanded with the deployment of machine learning systems. The aim of this work is to help provide solutions and ideas so that these groups have equal opportunities as the rest.

2 Programme and Methodology

The overall project will be done with two other people, who will be working on the English language, with one of them focusing on correference resolution and the other one on the hatespeech task. As for the workload distribution, the general pipeline will be set up as a group, while training the embeddings, the downstream task and the modification of the embedding space will be done mostly independently.

For our embedding algorithm, we will be using FastText, mentioned in section 1. While this is a relatively new embedding algorithm when compared to Glove and word2vec, it has seen increasingly more use. Another positive thing is that it generates the embeddings using subword information. This should help better capture the morphology of the Spanish language. Moreover, it can also help inform researchers dealing with more complex languages on whether reducing bias using this embedding algorithm is viable or not.

A quick overview of the workflow would be as follows. The embeddings will be trained using messages from Twitter and the FastText algorithm. Then, we will set up the hatespeech downstream task. This will use the hatEval dataset, mentioned in section 1. Then, we will set up our evaluation methods, both the WEAT test (intrinsic) and the Equality of Opportunity test (extrinsic). At this point, we will train the embeddings and the downstream task in order to be able to get a baseline with which to compare the test results once the embeddings have been modified.

We will use two different methods to reduce bias. The first one of these will be dataset balancing. One point to note is that the modifications done through this method are language and dataset-specific. Another important thing to note is that this method requires the embeddings to be retrained.

The second method we will use is the attract-repel algorithm. This algorithm is language-agnostic and does not require to retrain the embeddings. However, it does risk making changes that just fit the intrinsic metric, instead of actually changing the results on the downstream task.

Finally, we will compare the results of the modified spaces with those of the unmodified space. If we get satisfactory results and still have time left, we will use some of the other algorithms mentioned in section 1 to modify the embedding space, giving us better comparison points or use the other embedding algorithms mentioned in that section. This would in turn make our results more general. However, the main result of our work will not suffer if not enough time can be allocated for these extra tests.

2.1 Risk Assessment

Given the current world situation, sudden shifts in resources available could happen, which could lead to potential risks. The most notorious of these would be the lack of proper workspaces,

shifting of deadlines and, in a worst-case scenario, lack of the university's computing equipment.

The two largest issues would be the inability to control the embedding space and lack of access to computing equipment. The first case is a possible one, as our embedding space deals with subwords instead of whole words, which might lead to unexpected consequences, ranging from not being able to remove any bias at all to taking away any meaningful semantic representation. In this case, we would have to change the embedding algorithm that we're using. This would be major setback, but as the experiments are highly modular, it would just mean that less experiments will be able to be carried.

Finally, the lack of access to the university computing equipment would mean that we would have to turn to external providers, which could potentially limit our capacity to run experiments. However, we currently consider this a high risk but very low probability event and, in any given case, this situation should qualify for access to the recently announced covid emergency funding.

2.2 Ethics

No interaction with real subjects will be made at any point of the experiments. The datasets have already been generated by a third party and we will only have to comply with their terms and conditions.

3 Evaluation

We are trying to analyze the effect that altering the geometry of the embedding space has on bias on downstream tasks. As this has been observed to be reflected on the actual embeddings, we will be using and intrinsic and an extrinsic measures. These will be WEAT and Equality of Opportunity, respectively.

WEAT [17] is an intrinsic measure that mirrors association tests used with humans, such Harvard's implicit association test (IAT). These tests check how closely two concepts are associated, which can shine light into some unwanted associations that might appear. This measure is available in several languages like Spanish, Italian, and Russian, among others.

Equality of Opportunity [22] is an extrinsic measure that determines whether certain protected groups get treated differently under a certain task. The good thing about this measure is that its creators give guidelines on how to diminish discrimination when using this test without sacrificing much optimality.

4 Expected Outcomes

Given that there is no previous research on how altering bias on embedding spaces affects bias on downstream tasks, we are looking to find relations between the metrics used to measure this. Our expectation is that reducing bias on the embedding space will also lead to reduced bias on the downstream task. However, it is important to reiterate that our focus will reside in the relationship between the metrics rather than the numerical results.

5 Research Plan, Milestones and Deliverables

We will be dividing our tasks into three categories, namely, setup, experiments, and writing. A fourth category of stretch goals will be added for any extra experiments beyond those considered to be part of the minimum viable product. The relationships between these tasks and their expected duration can be seen in figure 1.

Setup

The result of each of these tasks is a python code that will form part of the pipeline. The tasks are:

(S_1) Implement the FastText embedding algorithm

(S_2) Implement the intrinsic evaluation method

(S_3) Implement the downstream task

(S_4) Implement the extrinsic evaluation method

(S_5) Implement the attract-repel algorithm

Experiments

The main results from these experiments are those coming from the metrics. We will also have embeddings from the embedding algorithms and the classification labels from the downstream task.

(E_1) Initial training of the embeddings

(E_2) Run experiment on the downstream task using the unaltered embeddings

(E_3) Do the balancing of the dataset and retrain the embeddings

(E_4) Use the attract-repel algorithm to get another set of altered embeddings

(E_5) Run experiments on the downstream task using both sets of altered embeddings

(E_6) Measure bias intrinsically and extrinsically on both the original and the altered embeddings

Writing

This part will just consist of writing the dissertation itself.

Stretch Goals

If we have enough time, we will work on two extra embedding algorithms: word2vec and GloVe. Their respective tasks would be the following:

- Implement the word2vec embedding algorithm

- Implement the GloVe embedding algorithm
- Initial training of the embeddings with both the original and the balanced dataset
- Use attract-repel to get a set of altered embeddings for each embedding algorithm
- Run experiments on the downstream task
- Measure bias intrinsically and extrinsically

Additional goals would be to implement extra methods to alter bias. Even though we most likely won't be able to reach these, the corresponding tasks for each additional method would be as follows:

- Implement the method used to alter bias
- Retrain the embeddings if needed
- Alter the embedding space if the method requires so
- Run experiments on the downstream task
- Measure bias intrinsically and extrinsically

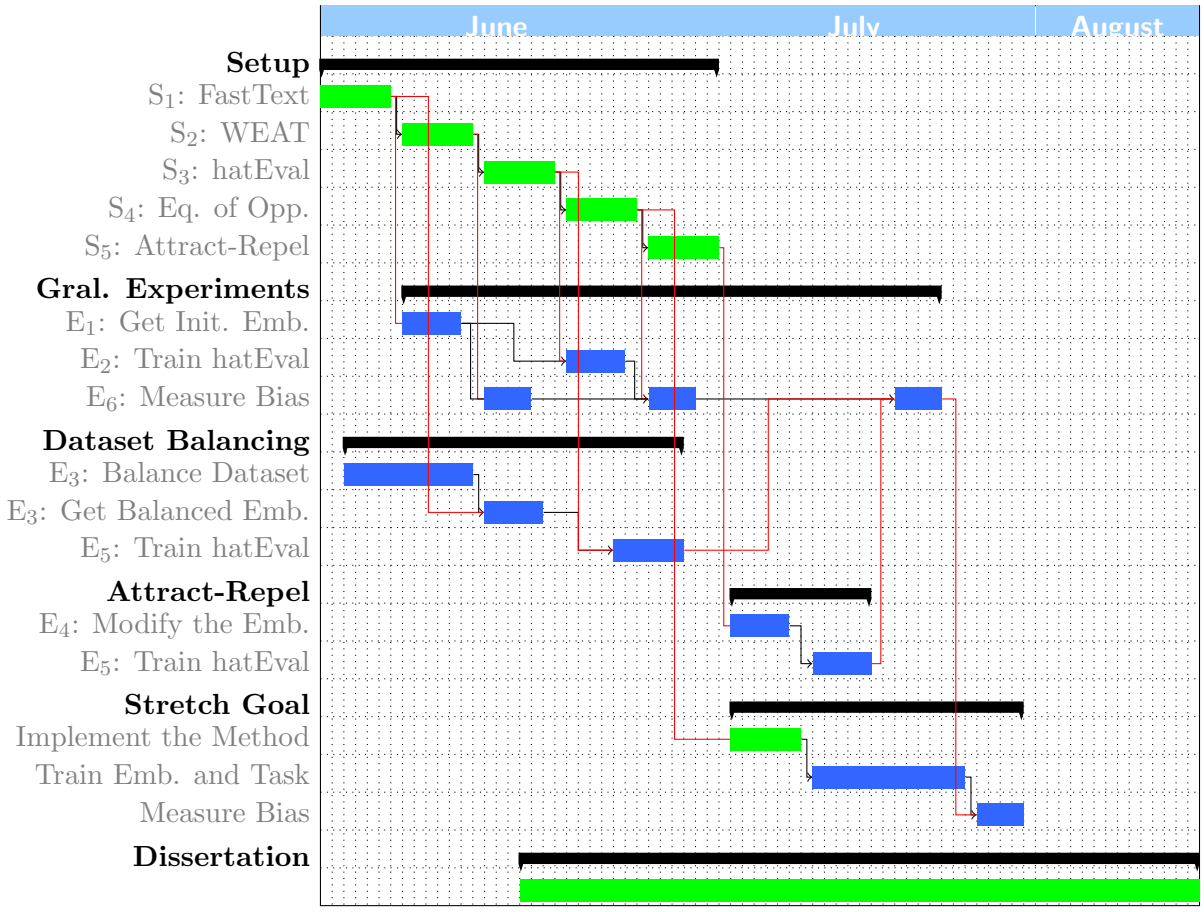

Figure 1: Gantt Chart for the activities described in section 5.

References

[1] Laurette Pretorius and Sonja E Bosch. Enabling Computer Interaction in the Indigenous Languages of South Africa: The Central Role of Computational Morphology. *Interactions*, 10:56–63, March 2003.

[2] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73, New Orleans, LA, USA, December 2018. Association for Computing Machinery.

[3] Zeerak Waseem. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics.

[4] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2016.00960.x.

[5] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, September 2013. arXiv: 1301.3781.

[7] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, December 2017. Publisher: MIT Press.

[9] Zhuang Bairong, Wang Wenbo, Li Zhiyu, Zheng Chonghui, and Takahiro Shinozaki. Comparative Analysis of Word Embedding Methods for DSTC6 End-to-End Conversation Modeling Track. In *DSTC6*, page 5, Long Beach, USA, December 2017.

[10] Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. Word Embeddings Go to Italy: A Comparison of Models and Training Datasets. In *IIR*, 2015.

[11] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364, Barcelona, Spain, December 2016. Curran Associates Inc.

[12] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2018.

[13] Hila Gonen and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *NAACL-HLT*, 2019.

[14] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[15] Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. How Does Grammatical Gender Affect Noun Representations in Gender-Marking Languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471, Hong Kong, China, November 2019. Association for Computational Linguistics.

[16] Katherine McCurdy. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. In *Anacode Wordpress Blog*, http://anacode.de/wordpress/wp-content/uploads/2017/06/mccurdy_grammatical_gender.pdf, June 2017.

[17] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. Publisher: American Association for the Advancement of Science Section: Reports.

[18] Latin American women prepare for record feminist marches. *Reuters*, March 2020. https://www.reuters.com/article/us-womens-day-latinamerica-idUSKBN20U0GE.

[19] Will Grant. Fury fuels historic women's strike in Mexico. *BBC News*, March 2020. https://www.bbc.com/news/world-latin-america-51736957.

[20] What is the migrant caravan heading to US? *BBC News*, November 2018. https://www.bbc.com/news/world-latin-america-45951782.

[21] Kirk Semple and Brent McDonald. Mexico Breaks Up a Migrant Caravan, Pleasing White House. *The New York Times*, January 2020. https://www.nytimes.com/2020/01/24/world/americas/migrant-caravan-mexico.html.

[22] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3323–3331, Barcelona, Spain, December 2016. Curran Associates Inc.