Polarizations in Static Word Embeddings: Investigating Bias in Marginalized Dialects of Portuguese

Raimundo Juracy Campos Ferro Junior
Universidade Federal do Ceará, Fortaleza, Brasil
ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), Portugal

Eugénio Ribeiro
INESC-ID Lisboa, Portugal
ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), Portugal

Leonardo Sampaio Rocha
Universidade Federal do Ceará, Fortaleza, Brasil

Abstract

This study investigates bias in static word embeddings applied to Pajubá, a dialect spoken within the Brazilian LGBTQIAP+ community. Four models from the NILC repository (Word2Vec, GloVe, FastText, and Wang2Vec) were evaluated across two dimensions: representational capacity and polarity analysis. The vocabulary coverage test revealed that approximately 74% of dialectal terms are present in the models' vector spaces. The RND and SC-WEAT tests consistently point toward negative associations for identity terms across all models, with the SC-WEAT further suggesting that reappropriated dialect terms carry their standard Portuguese positive valence into the embedding space, while explicit identitarian markers remain negatively encoded. An adjectivation test confirms stereotypical associations, including the linkage of *travesti* with criminality and the pathologisation of *gay*. These findings suggest that static word embeddings fail to capture the sociolinguistic complexity of Pajubá and systematically reflect the dominant and often discriminatory discourses present in the training corpora.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases word embeddings, bias detection, Portuguese, LGBTQIAP+, pajubá, NLP, lexical association, static embeddings

Digital Object Identifier 10.4230/OASIcs.SLATE.2026.7

Funding This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: `https://doi.org/10.54499/UID/50021/2025`), UID/PRR/50021/2025 (DOI: `https://doi.org/10.54499/UID/PRR/50021/2025`), and UID/04466/2025 (DOI: `https://doi.org/10.54499/UID/04466/2025`).
Raimundo Juracy Campos Ferro Junior: Funded by CAPES (PDSE scholarship)

1 Introduction

Text embedding models constitute the backbone of modern Natural Language Processing (NLP) systems. These algorithms transform linguistic units (texts, documents, sentences, expressions, and words) into continuous numerical vector representations that capture the semantic and syntactic relations of the source language.

Such embedding techniques enable algorithms, which are otherwise limited to numerical inputs, to interpret contextual similarities through mathematical operations. Although efficient for large-scale applications and widely adopted across diverse contexts, these models inherit structural patterns from their training corpora, encoding not only linguistic information but also latent **sociocultural biases** present in the data [4].

Investigating these biases reveals that discriminatory associations, such as gender stereotypes (e.g., *nurse* linked to women) and the negative racialization of certain terms, are systematically amplified in downstream applications (e.g., résumé screening, credit scoring). Several studies have sought to mitigate the effects of such biases [18, 15, 11, 9].

However, much of the research on bias in embedding models has focused on resources and techniques validated for English, rarely extending to Portuguese. There is therefore a significant gap in the literature regarding the evaluation of embeddings for Portuguese particularly across its dialectal and sociocultural varieties with a lack of analyses that account for the lexical and contextual specificities of the language.

This work addresses that gap. The primary objective is to evaluate the ability of embedding models to consistently represent the semantics of dialects belonging to marginalized LGBTQIAP+ communities. We hypothesize that, due to the limited presence of these dialects in widely used training corpora, terms characteristic of such linguistic varieties tend to be inadequately or entirely absent from the resulting embeddings.

We further investigate whether models can capture the contextual meanings of dialect words, both those shared with standard Portuguese and those exclusive to the dialect, as employed by their speakers. A salient example is the term *travesti* in the dialect Pajubá, where it functions as a neutral identity marker (carrying neither positive nor negative connotation) to address trans women, yet may be associated with negative connotations by the models due to biases present in the training data.

The specific objectives of this study are therefore: (1) to determine whether embedding models can generate vector representations for texts composed predominantly of marginalized dialect terms; and (2) to analyze whether such models reflect semantic biases by associating LGBTQIAP+ identity terms with negative or positive polarities through semantic polarity analysis.

2 Related Work

Bias detection in word embeddings has been an active research area since static embeddings were shown to encode stereotypical occupational associations [4], and WEAT was established as a standard intrinsic bias measurement tool [5].Despite the breadth of subsequent work, studies remain concentrated on English and on gender as the primary dimension of analysis, with sexual orientation, non-binary gender identities, and intersectional bias receiving considerably less attention [23].

The present work differs from this mainstream in two fundamental ways. First, it investigates a dimension of bias that remains underexplored: the association of LGBTQIAP+ vocabulary with negative attributes in static word embeddings. Second, it operates on a minority linguistic variety within Portuguese, Pajubá, rather than on standard register vocabulary.

Among the few studies addressing sexual orientation bias, [21] applied a WEAT-inspired metric using *straight* versus *gay* as target poles in German embeddings, finding that homosexuality-related terms clustered near concepts of violence, prohibition, and abuse, while heterosexuality was associated with positive sentiment. Our results for Portuguese show a similar directional pattern. However, their binary straight/gay opposition is insufficient for the broader spectrum of identities investigated here, which motivates the adoption of SC-WEAT in this work. [8] further showed, across fifteen English embedding models, that static embeddings systematically associate marginalised identities including LGBTQ groups with offensive vocabulary.

On the methodological side, [24] demonstrated that grammatical gender acts as a confounding signal in bias measurements for gendered languages, motivating the morphological adaptations made to the SC-WEAT attribute sets in this work. [7] provide the most directly comparable precedent, applying SC-WEAT to Portuguese GloVe embeddings and confirming that English bias metrics can be adapted to Portuguese with appropriate morphological adjustments.

3 Methodology

To achieve the proposed objectives, we adopted an approach grounded in tests consolidated in the literature. The investigation focused on a case study of Pajubá, a crypto-dialect developed and used by Brazilian LGBTQIAP+ communities, particularly by trans women. Further details on the dialect and the data collection process are provided in Section 3.1.

The experiments employed embedding models available in the **NILC** (Núcleo Interinstitucional de Linguística Computacional) word embedding repository.[14], covering four techniques: Word2Vec [19], GloVe [22], FastText [3], and Wang2Vec [17]. Details regarding the training corpus and model parameters are discussed in Section 3.3.

To address Objective 1 (representational capacity), we applied a **vocabulary coverage test**, which quantifies the proportion of dialect terms present in the models' vector space.

To address Objective 2 (polarity analysis), we employed three bias detection tests: **WEAT** (*Word Embedding Association Test*) [5], **RND** (*Relative Norm Distance*) [16], and the **Adjectivation Test** [5]. Each technique is defined and described in detail in Section 3.4.

3.1 Target Dialect: Pajubá

The marginalized community dialect selected for this investigation was Pajubá, also known as Bajubá. It is a crypto-dialect historically developed and used by Brazilian LGBTQIAP+ communities, particularly by trans women in peripheral urban contexts [1]. Its origins are tied to the need for safe intra-community communication [2]. The dialect has a cryptic character: everyday Portuguese terms are reappropriated with meanings known only within the community. Examples include *barbie* (to refer a specific body type of gay men), *Sida* (rearrangement of the AIDS acronym used to disguise the term. Even though *SIDA* is the Portuguese acronym for AIDS, in this dialect the term is repurposed as a feminine proper name, commonly used as *Aunt Sida*), and *Katia* (to refer to alcoholic beverages). This cryptic nature was not merely a linguistic curiosity, but a survival strategy: during the Brazilian military dictatorship (1964–1985), LGBTQIAP+ individuals faced systematic political persecution, driving the community to develop new vocabulary, particularly identity markers and terms referring to its own members, as a means of communicating safely in hostile environments. Consequently, Pajubá exhibits a significant concentration of terms with sexual connotations and gender and sexuality identity markers. These properties make Pajubá a particularly compelling case for stress-testing word embeddings: its reappropriated semantics, underrepresentation in standard corpora, and socially charged vocabulary collectively expose limitations that mainstream NLP benchmarks, typically anchored in standard language varieties, are unlikely to reveal.

3.2 Dataset

The dataset used in this study was drawn entirely from the *Dicionário Aurélia* [26], a Pajubá dictionary compiling expressions from the dialect. Entries were extracted from the digital edition of the dictionary using regular-expression-based text processing. A filtering step was

applied to retain only single-word terms, excluding multi-word expressions. This decision was made to simplify the analyses and to focus on individual lexical items. This process yielded a final corpus of 564 items, comprising a mixture of dialect-native terms, standard Portuguese words with altered meanings or polarities, and vocabulary derived from African languages. For the semantic polarity analyses (Objective 2), a manual categorization was performed to identify neutral terms, those designating people, entities, and objects within the sociocultural context of the speakers. Terms were selected based on three criteria: (i) functioning as a denotative referent within the dialect (i.e., terms used to refer to entities and objects in the speakers' context); (ii) intra-group semantic neutrality; and (iii) absence of intrinsic affective polarity (i.e., terms that do not convey evaluative judgment about the referent). For instance, the word *viado*, which outside the dialect is commonly used as a pejorative term, operates within Pajubá as a neutral noun or vocative among members of the gay community, and was therefore included in this selection, whereas terms with explicit affective charge were excluded, one example is *carimbada*, a derogatory term used to refer to people living with HIV. This subset of neutral terms comprises 159 words.

3.3 Embedding Models

We employed four static word embedding models from the NILC Word Embeddings Repository [14]: Word2Vec [19], GloVe [22], Wang2Vec [17], and FastText [3]. The NILC embeddings were trained on seventeen corpora of Brazilian and European Portuguese, totalling approximately 1.4 billion tokens from diverse sources including web text, news articles, literature, and technical content, with an emphasis on formal and standard registers, a characteristic particularly relevant to the underrepresentation of Pajubá discussed in subsequent sections. All models were used in their 100-dimension configuration; replication across other available dimensionalities constitutes a natural extension of this work.

3.4 Evaluation Methodology

Biases in embedding representations can manifest in various ways. Accordingly, we conduct several complementary tests, described below.

3.4.1 Vocabulary Coverage Test

The vocabulary coverage test was implemented to assess the ability of the models to represent Pajubá dialect terms, in line with Objective (i). This analysis consisted of verifying the existence of vectors associated with the 564 unique dialect terms in each embedding model. For each term in the dataset, a lookup operation was performed in the corresponding vector space, recording the presence or absence of a representation. The overall coverage rate was computed as the ratio of terms with a vector representation to the total number of terms in the dialect vocabulary.

3.4.2 Relative Norm Distance (RND)

The *Relative Norm Distance* (RND) metric [10] quantifies semantic bias by measuring the relative association between a set of target terms and two attribute groups representing opposite semantic poles. In this study, the attribute groups were defined as: (i) a set of 25 positive terms (e.g., *respeito*, *dignidade*, *bondade*) constituting the positive semantic pole, and (ii) a set of 25 negative terms (e.g., *maldade*, *ódio*, *crueldade*) constituting the negative semantic pole. The target set comprises Pajubá identity terms (e.g., *travesti*, *bicha*, *transformista*), extracted from the subset of 159 neutral terms described in Section 3.1, whose semantic polarity we wish to evaluate.

The RND computation begins by determining the centroid vector for each attribute group. The positive centroid ($\bar{v}_{T_1}$) is computed as the arithmetic mean of the vectors of all its terms; the negative centroid ($\bar{v}_{T_2}$) is obtained analogously. The vector difference between these centroids defines the bias axis onto which the target words are projected.

All vectors, both target terms and centroids, are subsequently ℓ_2-normalized to unit Euclidean norm, ensuring metric comparability and controlling for magnitude variation. For each target term a, its normalized projection onto the bias axis is then computed. The final RND score for the full attribute set corresponds to the arithmetic mean of these individual values.

Interpretation follows a straightforward principle: positive RND values indicate predominant association with the positive pole, while negative values indicate association with the negative pole. The absolute magnitude quantifies the strength of this association, with values near zero suggesting semantic neutrality.

In this study, the test was applied twice per model using two distinct attribute sets, both drawn from the 159-term neutral subset. The first set comprised 48 terms designating people and entities within the LGBTQIAP+ community; the second comprised 40 terms referring to objects and things. This distinction allows us to assess whether bias is specific to identity-bearing terms or extends more broadly across the dialect's vocabulary.

3.4.3 Adjectivation Test: Qualitative Analysis of Semantic Associations

The adjectivation test is a qualitative analysis that investigates the semantic associations between identity terms and adjectives in the vector space. For this test, five central terms of the LGBTQIAP+ community were selected: *travesti*, *transexual*, *gay*, *lésbica*, and *bissexual*.

The methodology involved three main steps. First, the **SentiLex-PT02** lexicon [6] was used as a reference resource. SentiLex-PT02 is a comprehensive inventory of Portuguese words annotated morphosyntactically and semantically, containing over 14,000 lemmas labelled with their grammatical class and affective polarity. All adjectives present in both SentiLex and the vocabularies of the embedding models under analysis were extracted, forming the candidate set for the association analysis.

Subsequently, for each target identity term, the cosine similarity between its embedding vector and all adjectives in the pre-filtered candidate set was computed, and the five closest adjectives were selected.

The objective of this test is to provide a characterisation of the semantic neighbourhood surrounding each identity term in the embedding space.

3.4.4 Word Embedding Association Test (WEAT) and SC-WEAT

The *Word Embedding Association Test* (WEAT), proposed by [5], measures semantic bias in word embeddings by testing whether two sets of target words (X and Y) are differentially associated with two sets of attribute words (A and B). Grounded in the *Implicit Association Test* (IAT) [12], WEAT treats cosine similarity between word vectors as analogous to reaction time in the IAT. For any word w, the differential association with attribute groups A and B is:

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(\vec{w}, \vec{a}) - \frac{1}{|B|} \sum_{b \in B} \cos(\vec{w}, \vec{b})$$

The aggregated difference across target groups and the standardized effect size d are then computed as:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \qquad d = \frac{\mu_X - \mu_Y}{\sigma}$$

where μ_X, μ_Y are the mean associations for each target group and σ is the pooled standard deviation. Statistical significance is assessed via a permutation test over all possible redistributions of $X \cup Y$.

However, WEAT requires X and Y to be of similar size and represent well-defined antagonistic poles. Some studies on sexual orientation bias construct these poles using terms such as *heterosexual* and *homosexual* [13], a valid approach for binary distinctions. The present work, however, investigates a broader spectrum of gender identities and sexualities for which no natural counterpart set exists, making such a bipolar structure incompatible with the problem studied.

This work therefore adopts the *Single-Category Word Embedding Association Test* (SC-WEAT), also proposed by [5] and applied to Portuguese embeddings by [7]. SC-WEAT operates on a single target set X, computing for each $w \in X$ the effect size:

$$\mathrm{ES}(\vec{w}, A, B) = \frac{\mathrm{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \mathrm{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\mathrm{std}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$$

Positive ES indicates association with pole A; negative with pole B. Significance is assessed via a one-sample permutation test based on sign rearrangements [13].

The attribute sets were constructed from the *pleasant vs. unpleasant* word lists of [5], adapted for Portuguese. Since WEAT was designed for English, a language without grammatical gender, direct application to Portuguese risks introducing morphological imbalances that distort results [24, 7]. To mitigate this, each gender-inflected adjective was included in both masculine and feminine forms, with its vector represented by the arithmetic mean of both. Morphologically invariant adjectives were retained in their single form. Additionally, the number of masculine and feminine adjective forms was balanced across both attribute sets, ensuring that no pole carries a disproportionate morphological gender signal that could confound the bias measurement.

4 Results

For the sake of a clear analysis, the results are organized according to the respective tests described previously.

4.1 Coverage of Dialectal Terms

The coverage tests demonstrated that only 74% of the terms present in the reference dataset were adequately represented in the analysed embeddings. Examination of the absent terms revealed that they consist predominantly of neologisms characteristic of the dialect under study and of terms originating from African-derived dialects, with particular note given to the significant absence of high-frequency expressions from *Pajubá*, such as *aqué* (money).

The terms that achieved adequate representation in the embeddings correspond primarily to vocabulary from standard Portuguese, with an emphasis on items of formal register and high frequency in standard corpora. This pattern suggests that the models capture the lexicon of written, institutionalised language far more reliably than the oral and community-specific vocabulary that characterises the dialect under investigation.

As an explanatory hypothesis for this phenomenon, the underrepresentation of dialectal markers follows directly from the low frequency of these terms in the written textual sources used to train the models. Since static word embedding models rely entirely on co-occurrence statistics derived from large written corpora lexical items that circulate primarily in oral, informal, or community-specific contexts are systematically marginalised in the resulting vector spaces.

4.2 RND (Relative Norm Distance) Test Results

Figure 1 presents the distribution of bias projections across the four embedding models for both word categories. Across all models and both categories, the median bias projection is negative, indicating a consistent tendency toward association with the negative attribute pole. This pattern is more pronounced in the People / Entities category, where 81% of projections are negative, compared to 72% in the Objects / Things category.

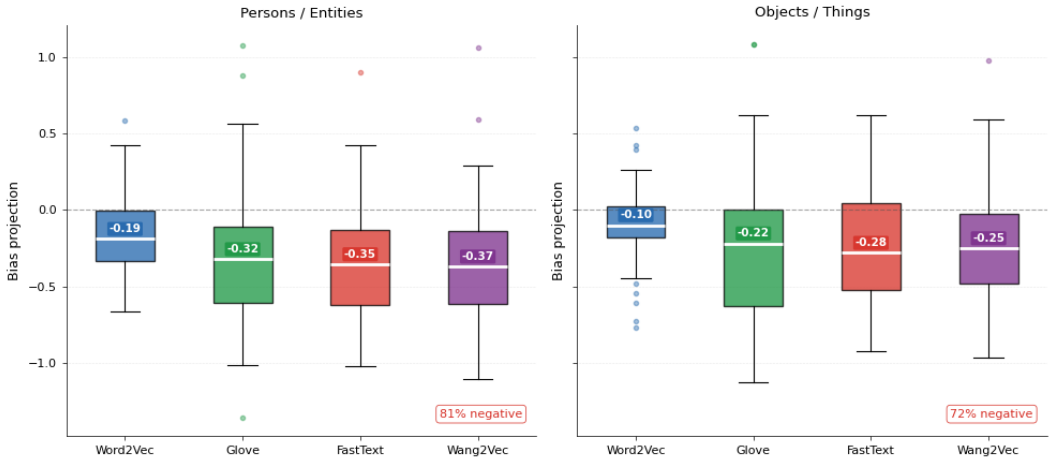

Figure 1 Distribution of bias projections across embedding models.

Among the models, Word2Vec exhibits the least negative median in both categories (−0.19 for People / Entities and −0.10 for Objects / Things), alongside the widest interquartile range, suggesting greater variance in its associations. GloVe, FastText, and Wang2Vec present more concentrated distributions with stronger negative medians, ranging from −0.32 to −0.37 in the People / Entities category and from −0.22 to −0.28 in Objects / Things. Notably, FastText and Wang2Vec display the most negative central tendencies overall, with medians of −0.35 and −0.37 respectively for People / Entities.

Outliers are present across all models in both categories, with isolated positive projections reaching values above 1.0, suggesting that a small subset of terms is associated with the positive attribute pole regardless of the model. Nevertheless, these exceptions do not alter the dominant negative trend observed across the full distributions.

The consistently negative projections across models and categories suggest that the target terms, spanning gender identities and sexualities, are systematically associated with negative attributes in the embedding spaces, regardless of the specific architecture or training corpus of each model. The stronger bias observed in the People / Entities category relative to Objects / Things may reflect that terms directly denoting people are more susceptible to encoding societal stereotypes present in the training data.

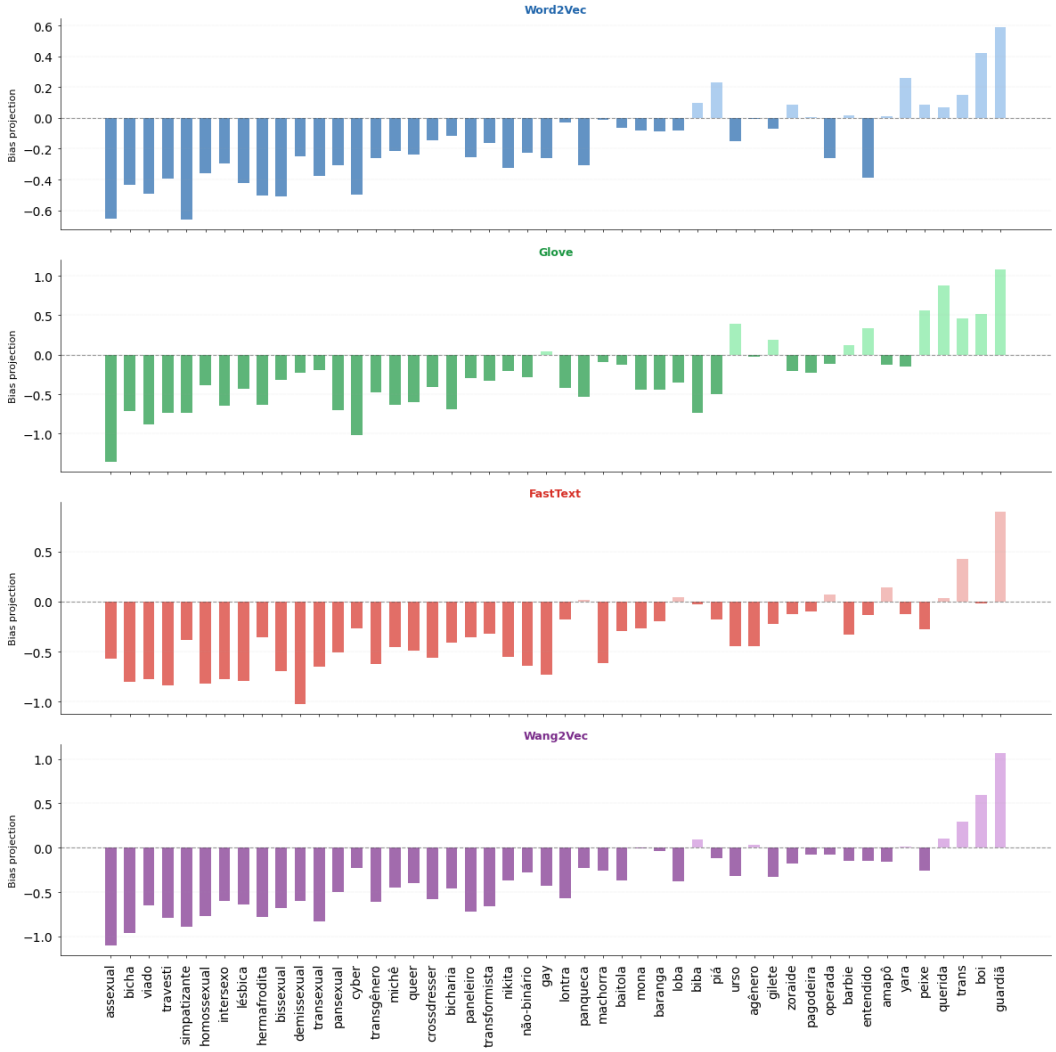

Figure 2 Bias projection – People / Entities.

Figure 2 presents the individual bias projections for each term in the People / Entities category across the four embedding models. The results reinforce the predominantly negative pattern observed in the aggregate analysis, with the majority of terms displaying negative projections across all models. However, a more granular examination reveals two observations.

First, while the direction of bias is largely consistent across models, the magnitude of individual projections varies considerably. Word2Vec tends to produce more moderate negative values, whereas GloVe, FastText, and Wang2Vec exhibit stronger negative projections for several terms, with some reaching values below -0.75 or even -1.0. This suggests that although all models encode a similar directional bias, the intensity with which individual terms are associated with negative attributes is sensitive to the architecture and training corpus of each model.

Second, the subset of terms presenting positive projections is consistent across models. Terms such as, *trans*, *boi*, and most notably *guardiã* recurrently appear with positive values across all four models. This cross-model consistency suggests that these associations are not artifacts of a specific model but rather reflect patterns systematically encoded in the Portuguese language corpora used for training.

A particularly notable case within this subset is *trans*, commonly used in Portuguese to refer to transgender individuals, and which consistently presents positive projections across all models. This stands in apparent contradiction with the strongly negative projections observed for related terms such as *transexual* and *transgênero*. However, this discrepancy is likely not a reflection of a more positive encoding of transgender identity per se, but rather a consequence of the morphological ambiguity of the term: *trans* functions as a productive prefix in Portuguese across a wide range of domains and its vector representation in the embedding space is therefore shaped by this broader and predominantly neutral or positive co-occurrence context, rather than by its community-specific usage alone.

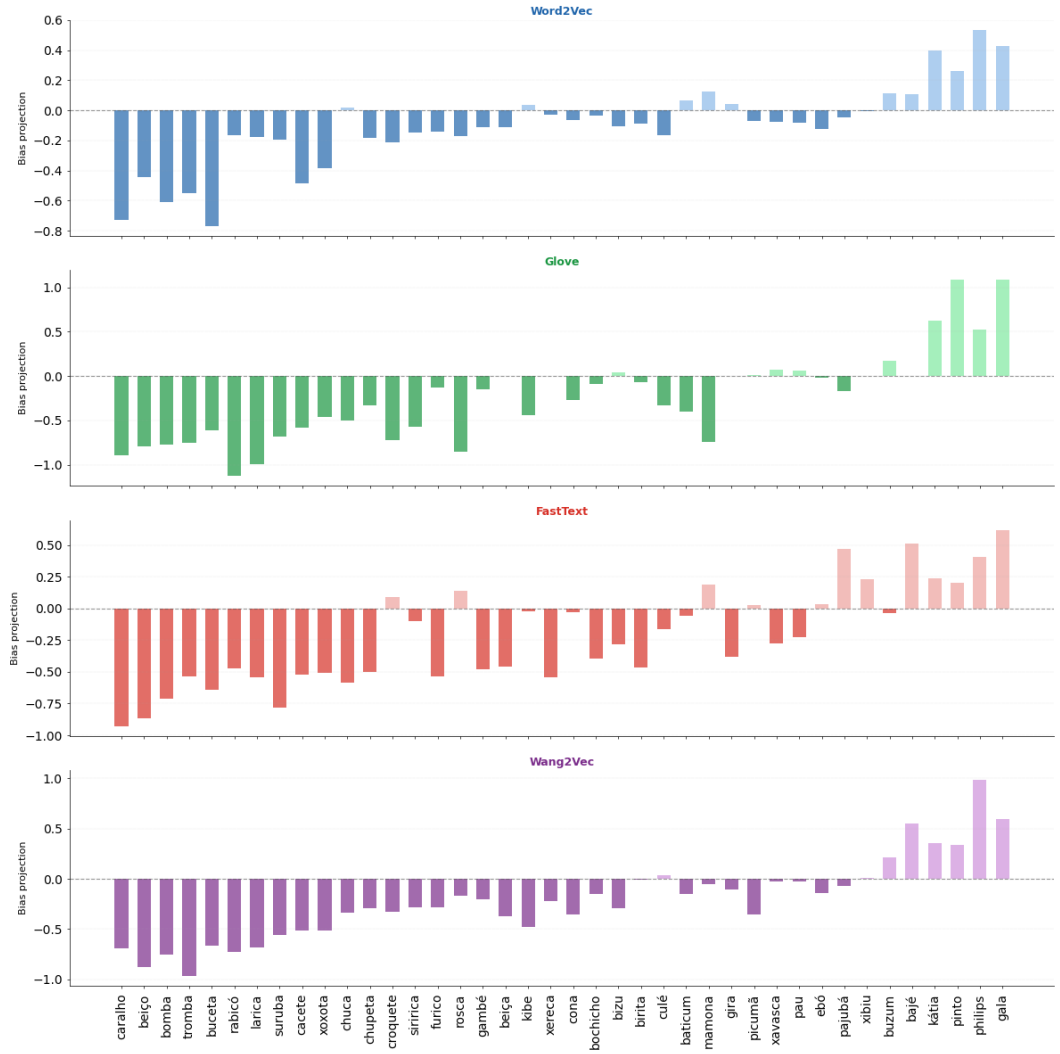

Figure 3 Bias projection – Objects / Things.

The Objects / Things category, presented in Figure 3, follows the same overall pattern, corroborating the 72% negative projection rate reported previously. The most strongly negative terms are concentrated at the left of the distribution and correspond to vocabulary with explicit sexual or anatomical connotations in Portuguese – such as *caralho*, *beico*, *bomba*, and *tromba* – whose pejorative or taboo framing is robustly captured across all models.

The positive subset comprises terms such as *philips*, *gala*, *pinto*, *kátia*, and *bajé*, several of which carry neutral or positive associations in standard Portuguese. As with *guardiã* in the previous category, their positive projections reflect formal Portuguese valence rather than community-internal reappropriation. The LGBTQIAP+ dialect operates across both ends of this spectrum, appropriating terms of positive valence while reclaiming terms of negative valence, but the models capture only the latter as negative, reflecting the dominant framing of these terms in the training data rather than their in-group meaning.

4.3 SC-WEAT Results

The SC-WEAT analysis was conducted in two stages. In the first stage, the same target terms from the RND analysis were used. It is important to note that although both methods investigate bias through the differential association of target terms with opposing attribute poles, they differ in how those poles are constructed: the RND approach uses general Portuguese positive and negative terms, while SC-WEAT contrasts a *pleasant* pole against an *unpleasant* pole based on the attribute word lists from [5], adapted for Portuguese as described in the methodology. The results of this first stage are presented in Table 1.

Table 1 SC-WEAT results for the RND target set.

Model	d	p	**Direction**
Word2Vec	+0.5202	0.5111	Pleasant
GloVe	−0.1789	0.7580	Unpleasant
FastText	+1.0067	0.2376	Pleasant
Wang2Vec	+0.4728	0.5121	Pleasant

None of the four models yielded statistically significant results, and three out of four presented a positive effect size, suggesting a tendency toward the pleasant pole. This outcome is unexpected given the predominantly negative bias observed in the RND analysis. However, this discrepancy is likely a consequence of the composition of the target set itself: as discussed in the previous section, the RND vocabulary includes a consistent subset of terms with positive valence in standard Portuguese, such as *guardiã*, *entendido*, and *trans*, whose presence in the target set may have attenuated or reversed the aggregate effect size, masking the negative associations observed at the individual term level.

To investigate whether this tendency toward the pleasant pole was driven by the standard Portuguese terms used to refer to the community, suggesting that these identitarian markers themselves carry positive associations, or by the reappropriated vocabulary of the dialect, a second stage was conducted using a refined target set restricted to explicit identitarian terms in standard Portuguese. If the identitarian markers consistently presented negative associations while the broader dialect vocabulary presented positive ones, this would constitute evidence that the cryptic function of the dialect is being fulfilled: the reappropriated terms successfully encode positive valence in the embedding space, while the bias against LGBTQIAP+ groups persists in the standard register. The target terms were divided into two subsets: gender-related terms (*travesti*, *transexual*, *trans*, *transgênero*, *não-binário*, *agênero*, *intersexo*, *queer*) and sexual orientation-related terms (*gay*, *lésbica*, *bissexual*, *homossexual*, *pansexual*, *assexual*, *demissexual*). The results are presented in Table 2.

The refined analysis reveals a more consistent pattern. Three out of four models present negative effect sizes for both subsets, indicating a tendency toward the unpleasant pole for terms explicitly associated with gender identity and sexual orientation. Wang2Vec

Table 2 SC-WEAT results for the refined target sets by category.

	Gender		Orientation	
Model	d	p	d	p
Word2Vec	−0.4955	0.6367	−0.3565	0.8129
GloVe	−0.9512	0.3028	−0.1817	0.8681
FastText	+0.1951	0.8638	+0.7346	0.7436
Wang2Vec	−1.3163	0.2001	−0.3883	0.8439

produces the strongest negative effect for the gender subset ($d = -1.3163$), followed by GloVe ($d = -0.9512$), while Word2Vec and FastText present more moderate values. FastText constitutes the sole exception, yielding positive effect sizes for both subsets, consistent with its behaviour in the first stage.

None of the results reach statistical significance, which warrants careful interpretation. The effect sizes observed, particularly for Wang2Vec and GloVe in the gender subset, are nonetheless substantial in magnitude and directionally consistent with the RND findings.

The persistent exception of FastText across both stages merits attention. Unlike the other three models, FastText represents words using subword information, decomposing terms into character n-grams before computing embeddings. This architecture makes it more robust to out-of-vocabulary terms but also means that the vector representation of a word is partially shaped by its morphological components. For terms such as *transexual*, *transgênero*, or *homossexual*, the subword structure may introduce associations with morphologically related but semantically distinct terms, potentially diluting or reversing the bias signal captured by the other models.

4.4 Adjectivation Test Results

The results of the lexical association test revealed consistent patterns of stigmatisation and identity reductionism across all models and terms analysed. For each of the five identity terms investigated (*travesti*, *transexual*, *bissexual*, *gay*, *lésbica*), two main phenomena were observed across all models: the recurrent association with other identity markers, and the attribution of pejorative terms. Figure 4 illustrates the word associations observed for each model.

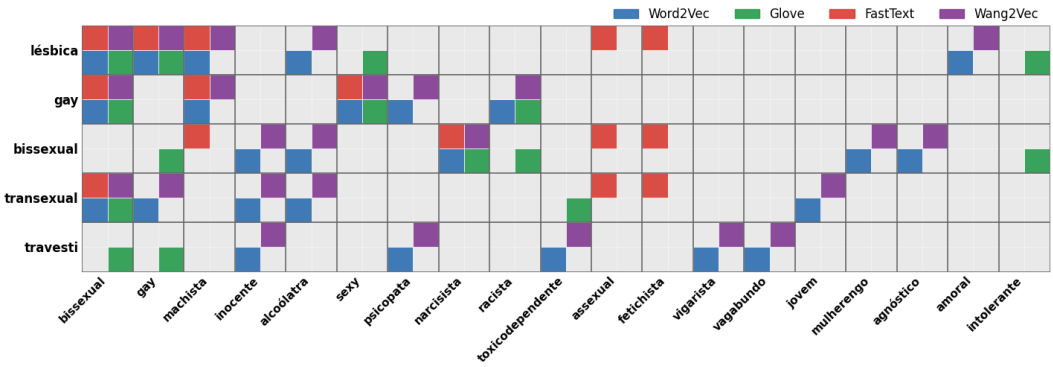

Figure 4 Adjectivation test results across all four embedding models.

The first phenomenon manifested through the conflation of distinct identities: *bissexual* was associated with *gay* across all models, and *lésbica* with *bissexual* in the majority of cases. This pattern suggests that the models fail to represent these identities as semantically independent, instead collapsing them into a shared neighbourhood of the embedding space.

The second phenomenon was evidenced by the systematic presence of negative attributes, including terms associated with criminality (*vigarista*, *bandido*, *delinquente*, *vagabundo*), pathologisation (*psicopata*, *alcoólatra*, *toxicodependente*, *maníaco-depressivo*), and negative moral judgement (*infiel*, *preconceituoso*, *amoral*, *intolerante*).

Certain term-specific patterns merit particular attention. The term *travesti* was associated with criminality-related expressions in three of the four models analysed. A plausible explanation for this pattern lies in the nature of the training data and the prevalence of media coverage of hate crimes against the trans population in the corpora. This discussion is elaborated in the following section.

5 Conclusion

The results obtained across the three analytical approaches consistently point to the presence of negative associations for vocabulary related to LGBTQIAP+ gender identities and sexual orientations in static Portuguese word embeddings, a pattern that persists across four architecturally distinct models and three complementary methods.

With respect to Objective 1, the vocabulary coverage test revealed that only approximately 74% of the dialect terms present in the primary source were found in the embedding models' vocabularies. This result indicates that static Portuguese word embeddings have limited representational capacity for the LGBTQIAP+ dialect, and that a substantial portion of this vocabulary remains inaccessible to embedding-based analysis. The terms that do appear are, by definition, those that have managed to cross the social boundaries of the speaker community and enter broader Portuguese language use. The terms that remain absent are likely those whose circulation is most strictly confined to in-group contexts, introducing a blind spot into any investigation that relies on these models.

With respect to Objective 2, the polarity analysis consistently pointed toward negative associations for the target vocabulary across all three methods and all four models. The models fail to capture the internal valence system of the LGBTQIAP+ dialect: terms that carry neutral polarity or that were reclaimed meaning within the community are not represented as such in the embedding space. Instead, the models reflect the dominant register of the training corpora, which frames these identities and their associated vocabulary through a predominantly negative lens. This is not a failure of the models per se, but an inherent consequence of training on data that disproportionately represents hegemonic discourse.

Other findings was that during the SC-WEAT, the aggregate tendency toward the pleasant pole appeared to contradict the RND findings. However, when the analysis was restricted to explicit identitarian terms in standard Portuguese, the negative tendency was restored across the majority of models. This divergence is interpretable as indirect evidence that the cryptic function of the dialect is partially reflected in the embedding space: reappropriated terms carry their standard Portuguese valence into the models, while the vocabulary used to explicitly name LGBTQIAP+ identities remains negatively encoded. It must be acknowledged, however, that the precise boundaries between positive and negative poles remain conceptually complex. The attribute word lists used to construct the SC-WEAT poles carry their own cultural assumptions, and the notion of what constitutes a pleasant or unpleasant association is not universally fixed. Refining these conceptual boundaries will be an important methodological challenge for future work in this area.

The adjectivation test adds a qualitative dimension to this picture. The recurrent co-occurrence of identity terms with vocabulary related to criminality, pathologisation, and moral condemnation is interpretable as evidence that these identities constitute a semantic pole in the embedding space, one that overlaps with these negative domains. A plausible hypothesis to account for this overlap lies in the social reality of LGBTQIAP+ communities in Brazil: the country consistently records the highest absolute number of lethal hate crimes against LGBT individuals in the world [20], and trans women are disproportionately relegated to sex work, a reality that is itself reflected in the dialect and in the contexts where this vocabulary circulates. Given that the training corpora include a substantial proportion of journalistic texts, the repeated reporting of violence, crime, and marginalisation involving these identities may have shaped their vector representations in the embedding space. It is therefore important to note that this does not imply that the models encode these identities as inherently criminal or pathological, but rather that these identities appear, in the training data, in contexts where such terms are present. The distinction between associative context and attributed characteristic is significant, and constitutes one of the most challenging interpretive dimensions of bias measurement in word embeddings.

5.1 Limitations and Future Work

This study is subject to several limitations that point toward directions for future work. The absence of approximately 24% of the dialect terms from the models' vocabularies represents a fundamental constraint, rooted in the scarcity of digitally accessible textual production by and about marginalised communities. Expanding the data sources through community-engaged collection or collaboration with LGBTQIAP+ organisations remains an important challenge. The small size of the SC-WEAT target sets further constrains statistical power, an inherent tension in analyses of minority linguistic communities.

The use of static embeddings, which assign a single vector to each word regardless of context, makes it impossible to distinguish between pejorative and reclaimed usages of the same term. Extending this analysis to contextual models such as BERTimbau [25] would allow for a more nuanced investigation of bias at the contextual level. The behaviour of FastText in the SC-WEAT analysis, consistently diverging from the negative tendency observed in the other models, may also suggest a bias dilution effect attributable to its subword architecture, a hypothesis that warrants dedicated investigation.

Finally, bias measurement in language models intersects with subjective, culturally situated questions of identity and valence that no single study can exhaust. As language evolves and the textual presence of LGBTQIAP+ communities grows, the biases encoded in future models may differ substantially from those observed here, making longitudinal follow-up a natural next step.

This study is intended as a first step toward a broader research agenda on bias in Portuguese language models for minority and marginalised linguistic communities, and we hope it encourages further investigation into the sociolinguistic dimensions of NLP fairness beyond the standard language register.

References

1 Karylleila dos Santos Andrade, Sheila de Carvalho P. Gonçalves, Filipe Porto, and Luciana C. e Silva Andrade. Bajubá: linguagem de grupo lgbtt como representação sócio-histórica e cultural. *Desafios – Revista Interdisciplinar da UFTO*, 5(4):36–46, 2018.

2 Gabriela Costa Araujo. *Bajubá: memórias e diálogos das travestis*. Paco Editorial, Jundiaí, SP, 1 edition, 2019.

3 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, 2017. doi:10.1162/tacl_a_00051.

4 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 *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357, 2016. URL: https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

5 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. doi:10.1126/science.aal4230.

6 Paula Carvalho, Mário J. Silva, Carlos Costa, and Luís Sarmento. Sentilex-pt: Principais características e potencialidades. *Oslo Studies in Language*, 7(1):e1444, 2015. doi:10.5617/osla.1444.

7 Fernanda Tiemi de Souza Taso, Valéria Quadros dos Reis, and Fábio Viduani Martinez. Analyzing discourses in Portuguese word embeddings: A case of gender bias outside the English-speaking world. *Journal on Interactive Systems*, 16(1), 2025. doi:10.5753/jis.2025.5958.

8 Fatma Elsafoury, Steven R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. SOS: Systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1263–1274, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. URL: https://aclanthology.org/2022.coling-1.108.

9 Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy, 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1166.

10 Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. doi:10.1073/pnas.1720347115.

11 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1061.

12 Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480, 1998. doi:10.1037/0022-3514.74.6.1464.

13 Bruce Pengfei Han. PsychWordVec: Word embedding research framework for psychological science. R package, 2025. URL: https://psychbruce.github.io/PsychWordVec/.

14 Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil, October 2017. Sociedade Brasileira de Computação. URL: https://aclanthology.org/W17-6615/.

15 Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/S18-2005.

16 Neeraja Kirtane and Tanvi Anand. Mitigating gender stereotypes in Hindi and Marathi. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 145–150, Seattle, Washington, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.gebnlp-1.16.

17 Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. Two/Too simple adaptations of Word2Vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi:10.3115/v1/N15-1142.

18 Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.naacl-main.76.

19 Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013. NIPS 2013. URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

20 Luiz Mott, Eduardo Michels, and Paulinho. Homicídios da população de lésbicas, gays, bissexuais, travestis, transexuais ou transgêneros (LGBT) no Brasil: uma análise espacial. *Ciência & Saúde Coletiva*, 25(6):2195–2208, 2020. doi:10.1590/1413-81232020256.19672019.

21 Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 446–457, 2020. doi:10.1145/3351095.3372843.

22 Jeffrey Pennington, Richard Socher, and Christopher D. 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, October 2014. Association for Computational Linguistics. doi:10.3115/v1/D14-1162.

23 Alexandre Puttick, Catherine Ikae, Carlotta Rigotti, Eduard Fosch-Villaronga, Mark W. Kharas, Roger A. Søraa, and Mascha Kurpicz-Briki. A systematic review of bias detection methods for non-English word embeddings and language models. *Artificial Intelligence Review*, 58:370, 2025. doi:10.1007/s10462-025-11375-8.

24 Saghar Omrani Sabbaghi and Aylin Caliskan. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 518–531, 2022. doi:10.1145/3514094.3534176.

25 Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems. BRACIS 2020*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer, 2020. doi:10.1007/978-3-030-61377-8_28.

26 Angelo Vip and Fred Libi. *Aurélia, a Dicionária da Língua Afiada.* Editora do Bispo, São Paulo, Brazil, 2006.