HeteroCorpus: A Corpus for Heteronormative Language Detection

Juan Vásquez
Posgrado en Ciencia e Ingeniería de la Computación
Universidad Nacional Autónoma de México
juanmv@comunidad.unam.mx

Gemma Bel-Enguix
Instituto de Ingeniería
Universidad Nacional Autónoma de México
gbele@iingen.unam.mx

Scott Thomas Andersen
Posgrado en Ciencia e Ingeniería de la Computación
Universidad Nacional Autónoma de México
stasen@comunidad.unam.mx

Sergio-Luis Ojeda-Trueba
Instituto de Ingeniería
Universidad Nacional Autónoma de México
sojedat@iingen.unam.mx

Abstract

In recent years, plenty of work has been done by the NLP community regarding gender bias detection and mitigation in language systems. Yet, to our knowledge, no one has focused on the difficult task of heteronormative language detection and mitigation. We consider this an urgent issue, since language technologies are growing increasingly present in the world and, as it has been proven by various studies, NLP systems with biases can create real-life adverse consequences for women, gender minorities and racial minorities and queer people. For these reasons, we propose and evaluate *HeteroCorpus*; a corpus created specifically for studying heterononormative language in English. Additionally, we propose a baseline set of classification experiments on our corpus, in order to show the performance of our corpus in classification tasks.

1 Introduction

In 1978, the french philosopher Monique Wittig gave a conference titled *The Straight Mind* (Wittig, 1979), in which she introduced the idea of the *straight regimen*. Wittig declared that heterosexuality is a political system that encompasses all aspects of western societies, and that its basis is the separation of people in binary and opposite categories based on their sex (Wittig, 1980). The author proposes that the idea of "women" –and that of all sexual minorities– is a generated byproduct of a "superior" category from which every institution should be modelled after. This category is, of course, "men" (Wittig, 1980).

Wittig also proposes that language is a system that has established that men, and heterosexuality, are the universals from which every particular derive from. This normalisation of heterosexuality as a political regimen through language –Wittig argues– contributes to the continuation of the oppressive systems against everyone who is not a member of the privileged "men" category (Wittig, 1980).

Adding to Wittig's ideas, Judith Butler proposed that *the subject is itself produced in and as a gendered matrix of relations* (Butler, 2011), meaning with this that the social and inner processes that construct the "subject" are deeply guided by the ideas of gender. Butler even remarks that the *matrix of gender* is generated prior to the creation of the subject, since this structure defines the limits and possibilities of what the subject can become (Butler, 2011). Therefore, the boundaries of what can be considered "human", are enforced by the matrix of gender, according to Butler.

Following these ideas, we hypothesize that the majority of the language used in current social media applications must exhibit numerous rules and expressions of heterosexuality as the norm.

In recent years, plenty of work has been done by the NLP community regarding gender bias detection and mitigation in language systems. Yet, to our knowledge, no one has focused on the difficult task of heteronormative language detection and mitigation. We consider this an urgent issue, since language technologies are growing increasingly present in the world and, as it has been proven by various studies, NLP systems with biases can create real-life adverse consequences for women, gender minorities and racial minorities.

For these reasons, we propose and evaluate *HeteroCorpus*; a corpus created specifically for studying heterononormative language in English. Our corpus consists of 7,265 tweets extracted from 2020 to 2022. In order to identify heterononormative lan-

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 225 - 234

guage in our corpus, we manually annotated every tweet, performed agreement experiments among the six annotators, and then evaluated the performance of our corpus in classification tasks using various classification systems.

The main contributions of our work are the following:

1. We present the first annotated corpus specialized in the study of heteronormative language.

2. We propose a baseline set of classification experiments on our corpus, in order to show the performance of our corpus in classification tasks.

The rest of the paper is structured as follows: Section 2 introduces the meaning of heteronormative and the negative impact it has had in society in general and the LGBTQIA+ community in particular. It also provides and overview of the work that has been done so far in gender bias detection and mitigation in NLP. Section 3 explains the configuration, annotation and challenges on compiling the HeteroCorpus, a data set especially designed for the detection of heteronormativity. In Section 4 we present the pre-processing and classification experiments. The results are discussed in Section 5. We close the paper with conclusions and future work (Section 6).

2 Related Work

In this section we will consider literature that explores what heteronormativity is and how the sense of the word has evolved over time, motivations to challenging heteronormativity, heteronormativity and gender bias as explored in natural language processing (NLP), and how this paper will contribute to this domain.

2.1 What is heteronormativity?

The word heteronormativity was coined by Warner (1991) and has been applied to a variety of contexts since then. The definition was recently analyzed and redefined to differentiate between these contexts (Marchia and Sommer, 2019). The authors propose formalizing the term heteronormativity to distinguish its usage among the following four distinct contexts; heterosexist-heteronormativity, gendered-heteronormativity, hegemonic-heteronormativity, and cisnormative-heteronormativity. We adapt the definition of heteronormativty from the dictionary CAER, (Diccionario de Asilo CAER-Euskadi), This definition translated to English is as follow:

> Heteronormativity refers to the social, political and economic regimen imparted by the patriarchy, extending itself through both the public and private domain. According to this regimen, the only acceptable and normal form to express sexual and affective desires, and even one's own identity is heterosexuality, which assumes that masculinity and femininity are substantially complementary with respect to desire. That is, sexual preferences as well as social roles and relationships that are established between individuals in society should be based in the 'masculine-feminine' binary, and always corresponds 'biological sex' with gender identity and the social responsibility assigned to it.

For simplicity, we seek to binarize the categorical definition of (Marchia and Sommer, 2019) this allows us to take advantage of binary decision classification of heteronormativity on our corpus.

Heteronormative speech has been found to create boundaries of normative sexual behavior, and relate to behaviors and feelings against violations of these norms. Results from recent investigation suggests that heteronormative attitudes and beliefs are relevant to political alignment and aspects of personality (Janice Habarth, 2015). Furthermore, we would like to bring to light The Gender Similarities Hypothesis, the idea that the biological sexes are more similar than they are different (Hyde, 2005). This is a stark contradiction to traditional arguments about biological differences between the sexes. Hyde finds that there is significant evidence to support her claim that many stereotypical biological differences between the sexes lack proper evidence to back them up, in fact, evidence seems to suggest the opposite in many cases. For example, some may believe that men are typically better than women at math, but Hyde's evidence concludes that the difference in mathematical ability is close to zero, and in some cases women outperform men.

Taking this into account with the claims of Habarth, we conclude that heteronormative speech has a substantial impact on perceptions of gender and sexuality, more so than actual biological differences between the sexes impact language.

2.2 Negative impact of heteronormativity

Given this definition we seek to justify the importance of detecting and challenging heteronormative ideology, not only to prevent harm but to promote gender equality and the inclusion of LGBTQIA+ people in society [1].

Recent investigation has shown that language can reflect sexist ideology. Coady (2017) has found that the process of iconisation, the partitioning of humans into two binary groups based on gender, can be projected onto language through sexist grammar and semantics in a process called fractal recursivity making the masculine gender the generic form. This linguistic gender norm leads to erasure of other genders and sexual identities from public discourse. Furthermore, Gay et al. (2018) demonstrate that presence of gender in language shows culturally acquired gender roles, and how these roles define house hold labor allocations. They go on to conclude that analysis of language use is promising because it is an observable and quantifiable indicator of values at the individual level These studies suggest that gender and sexual norms can be reflected in language use, Coady even concludes that the use of this language perpetuates such norms.

In fact, several recent studies have demonstrated that language use can be a subtle but effective barrier for gender minorities. Stout and Dasgupta (2011) demonstrate this by conducting experiments with mock job interviews with woman, finding that gender exclusive language during the interview negatively impacts the performance of women, however gender inclusive language, i.e. "he or she", or gender neutral language, i.e. "one", led to an improved performance among women. Meanwhile Davis and Reynolds (2018). demonstrate that using language that normalizes the binary sex classification is strongly associated with a gender gap in educational attainment. That is, heteronormative language is not only indicative of sexual and gender disparity, it also is a proponent of it.

Research shows that not only does heteronormative speech disadvantage women, patterns in language use on social media can be indicative of psycho-social variables demonstrating personality traits and emotional stability among men and women. For example, men more commonly use possessive pronouns before nouns referring to a female partner, i.e. "my girlfriend" (Schwartz et al., 2013). Eaton and Matamala (2014) even find that heteronormative beliefs about men and women may encourage sexually coercive behavior in intimate relationships.

Many of these previous studies have dealt with language use and it's relationship with discrimination based on the "men and women" gender binary. Let us know to explore research on heteronormative language and it's effect on LGBTQIA+ individuals. Lamont (2017) finds in a survey of LGBTQIA+ individuals, that the majority report finding that the heternormative script of relationships are constraining, unimaginative, and heavily gendered, suggesting that many members of the queer community feel restricted by the expectation set by heteronormative values. While Smits et al. (2020) analyzed heteronormative speech and casual use of homonegative slurs in young men in sports and found that this language was used almost devoid of meaning except to express lack of masculinity, disapproval, and negativity, concluding that this use of speech attributes to the preservation of heteronormative discourse in spite of growing acceptance of non-heterosexual male athletes. Another study finds that many LGBTQIA+ social work students experience an overwhelming amount of discrimination, mostly perpetuated through harmful discourse (Atteberry-Ash et al., 2019). Lastly, King (2016) finds that heteronormative speech and policing of gender roles in children lead to hypermasculine and violent men, concluding that violence to the queer community can all be connected to heteronormativity in everyday life.

2.3 Gender bias detection and mitigation in NLP

While heteronormativity refers to a more comprehensive system, gender bias is an element to this system since both are based on the idea of creating separate realities for people according to one of the two genders they were assigned at birth. Since, to the best of our knowledge, there is no literature on heteronormative language detection in NLP systems, we choose gender bias efforts as both motivation and justifaction for our work. Gender bias is the preferential treatment towards men over women, often unintentionally and exhibited by all

[1] Here we wish to clarify that we promote preventative action against all gender and sexual discrimination. LGBTQIA+ refers to the lesbian, gay, bisexual, transgender, queer, intersex, asexual communities as well as all additional gender and sexual identities that deviate from the traditional heteronormative relationship.

genders (Corinne A. Moss-Racusin et al., 2012).

To continue, we will take a look recent literature that seeks to address gender bias in the NLP space. Sun et al. (2019) address this with a literature review, bringing to light the lack of research pertaining to gender bias in NLP, and a lack of concrete methods for detecting and quantifying gender bias. They go on to address that debiasing methods in NLP are frequently insufficient for end-to-end models in many applications. We envision our corpus contributing to the development and verification of methods for the detection of that arises from heteronormative language.

Recent work has come forth to formalize how gender should be considered ethically in the development (Larson, 2017), bringing to light how many recent studies have brought gender as a variable in their experiments whilst assuming binary categories. Most often however, it was found that many recent or widely cited papers gave little to no explanation for how they defined these categories, simply describing the variable as "gender" or "sex" without further clarification. This is indicative of a heteronormative mindset used in much of NLP research.

The bias of researchers can be reflected in the work they are doing, and we hope that the work that comes from our anti-heternormative dataset can bring these biases to light.

Lu et al. (2018) propose a metric to quantify gender bias in NLP in response to existing models that exhibit bias, such as text auto-completion that makes suggestions based on the gender binary. They also propose a method to mitigate gender bias. Bordia and Bowman (2019) address existing language models and point out the gender bias that they contain. They note that many text corpora exhibit problematic biases that an NLP model may learn. Gender bias, as we have seen, can reflect and be perpetuated by heteronormativity. However, the scope of our work is to further generalize the bias in question to go beyond the gender binary and include LGBTQIA+ people. Dev et al. (2021) survey non-binary people in AI to illustrate negative experiences they have experienced with natural language systems. They challenge how gender is represented in NLP systems and question whether we should be representing Gender as a discrete category at all.

Once the NLP community established that gender biases indeed exist in many NLP systems, many efforts have been made towards detecting and mitigating these biases. Next, we mention some of these techniques in various NLP tasks and systems: from machine translation, coreference resolution, word embeddings, large language models to sentiment analysis. First, we focus on the works regarding large language models, specifically, BERT. Bhardwaj et al. (2020) state that contextual language models are prone to learn intrinsic gender-bias from data. They find that BERT shows a significant dependence when predicting on gender-particular words and phrases, they claim such biases could be reduced by removing gender specific words from the word embedding. Zhao et al. (2018) go on to produce gender-neutral word embeddings that aim to preserve gender information in certain dimensions of word vectors while freeing others of gender influence, they release a gender neutral variant of GloVe, GN-GloVe. Kurita et al. (2019) proposes a method to measure bias in BERT, which successfully identifies gender bias in BERT and exposes stereotypes embedded in the model. Recent models have been developed to mitigate gender bias in trained models, such as Saunders and Byrne (2020), who use transfer learning on a small set of gender-balanced data points from a data set to learn un-biasedly, rather than creating a balanced dataset.

Many recent efforts focus on the creation of corpora for gender bias detection and mitigation. Such as Doughman and Khreich (2022), who create a text corpus avoiding gender bias in English, much like our research, however we focus on heteronormativity. Likewise, Bhaskaran and Bhallamudi (2019) create a dataset that is used for detecting occupational gender stereotypes in sentiment analysis systems. Parasurama and Sedoc (2021) state that there are few resources for conversational systems that contain gender inclusive language. Cao and Daumé III (2020) present two data sets. `GAP` which substitues gender indicative language for more gender inclusive words, such as changing *he* or *she* for the word *they* or neopronouns. They also present `GIcoref`, an annotated dataset about trans people created by trans people.

Finally, we mention two works focused on gender-neutral pronouns in NLP systems. We find these efforts relevant to our work, since a way to challenge heteronormative language is to eliminate the gender markers in language altogether. Lauscher et al. (2022) provide an overview for gen-

der neutral pronoun issues for NLP, they propose when and how to model pronouns, and present demonstrate that the omission of these pronouns in NLP systems contributes to the marginalization of underrepresented groups. Finally, Bartl et al. (2020) studies gender bias in contextualized word embeddings for NLP systems, they propose a method for measuring bias in these embeddings for English.

These systems deal typically with detection and identification of gender bias. Research that attempts to include gender minorities deals with the issue of a lack of resources that can identify bias from heteronormativity. This paper aims to solve that problem by providing a dataset that can use existing debiasing techniques to address bias that stems from heteronormativity.

3 HeteroCorpus

In this section we will describe our process for collecting data from Twitter and the annotation process, as well as the challenges we faced and the resulting dataset.

3.1 Data Statement

We follow the guidelines specified by (Bender and Friedman, 2018) to produce a Long Form data statement. A data statement is important when producing NLP datasets to mitigate bias in data collection.

A. Curation Rationale We collect tweets from popular social media platform Twitter, we use Twitter because it provides a convenient medium to collect short statements from general users in on various topics in a digital medium. We use specific search terms that are indicative of gender because we aim to build a dataset that consists of heteronormative speech.

B. Language variety We scrapped a set of tweets that contained desired keywords and were in English. However, there were tweets present in other languages, and we instructed annotators to indicate them using a separate tag so they could be discarded. There are no restrictions on the region from which the tweet could come. Since all the data is collected from social media, this means the presence of hashtags, mentions, gifs, videos, images, and emojis within the tweets. Also, we found spelling mistakes, abbreviations and slang native to social media.

C. Tweet author demographic The demographics of the authors is not available to us since we compiled the data by the tag `EN` that Twitter provides; however, due to our sampling methods, we expect the tweets to come from a diverse set of authors of various ages, genders, nationalities, races and ethnicities, native languages, socioeconomic classes and education backgrounds.

D. Annotator demographic All the annotators are students members of Grupo de Ingeniería Lingüística (Language Engineering Lab) from the Universidad Nacional Autónoma de México. The demographic information is shown in 1.

Categories	**Data**
Age	20-25 years
Gender	3 women 3 men
Sex	3 female 3 male
Sexual Orientation	2 Heterosexual 2 Homosexual 1 Bisexual 1 Demisexual
Nationality	5 Mexican 1 American
Residence	6 Mexico
Field of Study	3 Linguistics 1 English Literature 1 Translation 1 Computer Science
Native Language	5 Spanish 1 English
Secondary Language	5 English 1 Spanish

Table 1: Demographics as anonymously self reported by each annotator.

E. Speech Situation Each tweet may have a different speech situation. Most of them are related to tendencies, events or memes from the year of extraction (2022).

F. Text characteristics The tweets collected come from a diverse set of contexts, as they could be published alone by the author, or in response to another user. The tweets are subject to the restrictions of text limit and policies of Twitter. All tweets were posted publicly, and we remove identifying characteristics of the user for anonymity.

G. Recording Quality We extracted the tweets from the Twitter API.

3.2 Data Collection

The first step was to acquire a set of tweets that could potentially contain heteronormative language used by the authors. To do this we crafted a list of terms that we noticed had several heavily gendered trends while reading tweets. These terms are the following: *man, men, husband, son, boy, woman, women, wife, daughter, girl*. In this selection, we have tried to avoid heavily-gendered and queer terms, to focus in the most general framework. However, we are aware that this can introduce bias.

After defining the terms for our search, we performed the extraction of the tweets via the Twitter API. For each term, specifically in the English language, we performed a search for the period of time ranging from 1 Jan. 2020 to 10 Mar. 2022. The total number of extracted tweets was 26,183.

The next step was to perform a filtering of the obtained tweets. The first filter was based on the presence or absence of adjectives in the tweets. First, we obtained a list of the adjectives in the entire dataset. Then we used that list to create another list with terms that followed the syntactic structure: `adjective` + `relevant search term` or `relevant search term` + `adjective`. For example, we found the adjective *nice* among the tweets crawled. Therefore, all the tweets with the pairs *nice man, girl nice, etc* were kept for the next stage of filtering, since they contained a relevant search term and an adjective. The motivation behind this filter was that, by manually observing the crawled tweets, we noticed that those tweets with the syntactic structure described above contained some of the most heteronormative discourses in them. This made sense for us since it is well known that the use of adjectives in English has reflected gender bias (Rubini and Menegatti, 2014).

After the first filter, we obtained a dataset with 9,350 tweets in it. From those tweets, we removed those that only contained our search terms. For example, tweets with only the text "man!" were removed. We decided to do this because we considered that those tweets did not contain a great amount of semantic information relevant to heteronormative language, and were only indicative of a conversation having place.

The final size of our dataset was 7,265 tweets. The frequency distribution of the terms in our final corpus is shown on Table 2.

Term	Frequency	Term	Frequency
man	3070	woman	1713
men	1285	women	33
husband	708	girl	1056
boy	844	wife	740
son	655	daughter	1072

Table 2: Number of times each of the key terms appears in the HeteroCorpus.

3.3 Annotation Protocol and Results of the Annotation Process

The first step in the creation of the annotation protocol, was to establish the two labels that could be assigned to the tweets. These labels were *0 - Non-Heternormative* and *1 - Heteronormative*. We also gave the annotators the option to set a label *2* for the tweets that did not have any content relevant to the topic of the corpus. Some tweets labeled with *2* were those that only contained hashgtags (#) or mentions (@). Also, the tweets in other languages and those with only emojis in them were assigned a label of *2*. The tweets under this class were removed once the annotation was finished.

Afterwards, we wrote the Annotation Guide[2], in which we defined what the annotators should understand as *heteronormativity* 2.1. Furthermore, we randomly selected a sample of 100 tweets, and assigned a copy of this subset to each annotator before beginning the final annotation process. Each annotator was provided with their own Google Drive Spreadsheet document that contained the following four columns: the number of the tweet, the tweet, the ID, and the label. We asked the six annotators to classify the tweets in this test sample.

Then, we organized a meeting with the annotators in order to test how this annotation process turned out. In that meeting, the authors of this paper evaluated the performance of each annotator. We asked them to justify various label decisions they made and their thought-processes behind their annotations. Then, we gave them all some feedback on their annotations. Finally, we all discussed how to settle ambiguous cases.

[2]This annotation guide is available in the GitHub with the HeterCorpus dataset.

The next step in the annotation process was the annotation of the entire dataset. From the 7,265 tweets that comprised our dataset, we shuffled them randomly and split them in two partitions. The first partition had a size of 3,632, while the latter one had a size of 3,633. Three annotators were assigned to work on the first partition, while the other three annotators worked on the second one. In total, each tweet was annotated three times.

Once the annotators were done, we obtained Cohen's Kappa on the annotation pairs. Using these calculations, we set on the final labels for each tweet. The pairs with an agreement of 3/0 or 0/3 made up 65% of the dataset, while the pairs with an agreement of 2/1 or 1/2 constituted the remaining 35% of the tweets. We also obtained the Fleiss' Kappa on the entire dataset. The value of this calculation was 0.4036. The final distribution of the labels was of 5,284 tweets with the label 0, and 1,981 tweets with the label 1.

A few examples of tweets can be found in Table 3.

4 Methodology for Heteronormativity Detection

In order to establish a baseline for classification systems trained on our corpus, we performed a set of classification experiments.

4.1 Data Pre-Processing

First, we removed the urls in the dataset. Then, we tokenized and lemmatized our entire corpus. Afterwards, we removed the mentions, punctuation marks, and stop-words[3].

The next step was to create the training and evaluation sets. For this, we split the corpus into two partitions: the first one, with 90% of the tweets in the original corpus, and the second with the remaining 10% tweets.

4.2 Classification Experiments

After the text pre-processing steps, we implemented two supervised classification algorithms. The first, a SVM classifier using as features a combination of bag-of-words with TF-IDF[4], the second was performed using a logistic regression algorithm. For both steps, we used the same features as previously described.

Various works have focused on sexism classification in English (Jha and Mamidi (2017), Bhaskaran and Bhallamudi (2019)). In order to have a starting point for our experiments, we followed their steps with the use of SVM and logistic regression algorithms.

Afterwards, we proceeded to test our corpus on a binary classification task using deep-learning architectures; specifically, four different versions of BERT, following de Paula et al. (2021)'s work. These authors obtained the highest accuracy and F1-score on a sexism prediction shared task organized on 2021 at the IberLef 2021 using a corpus comprised of tweets in English and Spanish.

We fine-tuned the BERT-base-cased, BERT-base-uncased, BERT-large-cased, and BERT-large-uncased models[5]. The hyperparameters used while fine-tuning the BERT models were the following, as suggested by the original authors of BERT (Devlin et al., 2018). We use 4 epochs, and a batchsize of 8; the learning rate is $2e^{-5}$ with $1e^{-8}$ steps and a max sequence length of 100 tokens. Finally, we use the `AdamW` optimizer.

5 Results and Discussion

Since the task of identifying heteronormativity in NLP systems has not been studied yet, we compare our classification experiments with systems that detected gender bias. We decided not to compare with hate speech tasks, since we consider that heteronormative language does not necessarily imply hate speech.

We recognize that our baseline can only be vaguely compared with the results obtained by other authors in other classification tasks, since we aim to detect different linguistic phenomena. Following those remarks, on Table 4 we show the results obtained on our heteronormativity detection experiments.

It can be observed that BERT-large outperforms the supervised classification algorithms. Also, the low results shown on Table 4, indicate that the task of classifying heteronormativity is not a simple one and more work will be required in order to improve the results of this benchmark.

6 Conclusion and Future Work

In this paper, we presente HeteroCorpus; a novel human-annotated corpus for heteronormative language detection. This work sets a new precedent

[3]For this we used the pre-loaded set of stop-words in English provided by nltk

[4]The implementation of TF-IDF we used was the one provided by the scikit-learn library.

[5]We implemented scikit-learn's wrapper for BERT.

Tweet Text	Label
Your life, little girl, is an empty page that men will want to write on	1
This is utter bullshit, plenty of women find heavier set men attractive.	1
ur boy could most definitely use a friend this week.	0
Sweet man! Yeah, it took a minute but I'm glad I didn't have to buy from resellers	0
Beautiful you filmpje Geil beautiful you lull I your broekje are very beautiful man [...]	2

Table 3: Example tweets from the HeteroCorp. Here we present some examples of tweets, their categorization, and the reviewer agreement. 1 indicates the tweet is heteronormative, and 0 indicates the tweet is non-heteronormative. 2 indicates a tweet that was in another language or was not intelligible.

Classifier	Accuracy	F1-score
SVM	0.64	0.55
LR	0.67	0.50
BERT-base-uncased	0.63	0.59
BERT-base-cased	0.68	0.62
BERT-large-uncased	**0.71**	**0.72**
BERT-large-cased	**0.72**	**0.72**

Table 4: Results for the heteronormativity detection experiments using our corpus.

in NLP, since, to the best of our knowledge, there has not yet been developed a similar corpus that aims to study heteronormative language in English. We consider that this corpus could be of use in gender bias and sexism detection and mitigation tasks, which have proven to be quite challenging. While gender bias and sexism are not the same as the presence of heteronormativity in language, they all are nocive issues present in current NLP systems. Until the NLP community finds an efficient way to minimize these issues, language technologies will continue to amplify the discrimination based gender and sexual identity.

The Fleiss' Kappa obtained on our corpus signals a moderate agreement between our annotators. This indicates that annotating heteronormativity can be complicated. Therefore, researchers must take into consideration this extra challenge while creating similar resources, since the quality of the data depends on the expertise of the annotators.

We also present a baseline for the task of heteronormative language detection using our corpus, with two supervised algorithms and with four variations of BERT.

As future work, we plan on expanding this corpus by extracting a larger set of tweets containing more nuanced forms of heteronormative discourses, since heteronormativity is not only associated to lexical properties in the speech, but also to more complex forms of linguistic phenomena. In future projects, we hope to further investigate heteronormative language use in digital spaces, crafting a dataset that better respects the multi-class definition of heterornormativity as discussed in Section 2.

We propose the creation of similar corpora but for other languages, since heteronormativity is a global issue that requires joint action. Also, we encourage researchers to develop further tools for heteronormative language detection and mitigation, since language technologies are rapidly increasing their presence in human lives, and the implicit biases these models have can be very costly and damaging to human lives.

7 Ethical Considerations

7.1 Data Collection

We ensured that our dataset was obtained following Twitter's terms and conditions.

The full text corpus will not be released due to Twitter's Privacy Policy. Only the IDs of the tweets and their labels are be available on the following repository[6].

7.2 Benefits and Limitations in the use of our Data

This corpus has been created for the detection of heteronormative language in English. Other possible uses could be gender bias and sexism detection and mitigation. Every population could be benefited from the integration of our corpus into their language systems, since it's main goal is to create more equal language technologies.

8 Acknowledgements

This paper has been supported by PAPIIT project TA400121, and CONACYT CB A1-S-27780. The

[6]https://github.com/juanmvsa/HeteroCorpus

authors thank CONACYT for the computing resources provided through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo del INAOE

References

Brittanie Atteberry-Ash, Stephanie Rachel Speer, Shanna K. Kattari, and M. Killian Kinney. 2019. Does it get better? LGBTQ social work students and experiences with harmful discourse. *J. Gay Lesbian Soc. Serv.*

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *CoRR*, abs/2010.14534.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2020. Investigating gender bias in BERT.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. *arXiv preprint arXiv:1906.10256*.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *NAACL*.

Judith Butler. 2011. *Bodies that matter: On the discursive limits of sex*. routledge.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Ann Coady. 2017. The origin of sexism in language. *Gender and Language*.

Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479.

Lewis Davis and Megan Reynolds. 2018. Gendered language and the educational gender gap. *Econ. Lett.*

Angel Felipe Magnossão de Paula, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. *arXiv preprint arXiv:2111.04551*.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diccionario de Asilo CAER-Euskadi. Heteronormatividad género y asilo. Online. Accessed: 2022-04-06.

Jad Doughman and Wael Khreich. 2022. Gender bias in text: Labeled datasets and lexicons.

Asia A. Eaton and Alejandra Matamala. 2014. The relationship between heteronormative beliefs and verbal sexual coercion in college students. *Arch. Sex. Behav.*

Victor Gay, Daniel L. Hicks, Estefania Santacreu-Vasut, and Amir Shoham. 2018. Decomposing culture: an analysis of gender, language, and labor supply in the household. *Review of Economics of the Household*.

Janet Shibley Hyde. 2005. The gender similarities hypothesis. *Am. Psychol.*

Janice Habarth. 2015. Development of the heteronormative attitudes and beliefs scale. *Psychology and Sexuality*.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Jessica King. 2016. The violence of heteronormative language towards the queer community.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations.

Ellen Lamont. 2017. "we can write the scripts ourselves": Queer challenges to heteronormative courtship practices:. *Gend. Soc.*

Brian Larson. 2017. Gender as a variable in Natural-Language processing: Ethical considerations. *EthNLP@EACL*.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-Inclusive natural language processing beyond gender.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv: Computation and Language*.

Joseph Marchia and Jamie M. Sommer. 2019. (re)defining heteronormativity. *Sexualities*.

Prasanna Parasurama and João Sedoc. 2021. Gendered language in resumes and its implications for algorithmic bias in hiring.

Monica Rubini and Michela Menegatti. 2014. Hindering women's careers in academia: Gender linguistic bias in personnel selection. *Journal of Language and Social Psychology*, 33(6):632–650.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The Open-Vocabulary approach. *PLoS One*.

Froukje Smits, Annelies Knoppers, and Agnes Elling-Machartzki. 2020. 'everything is said with a smile': Homonegative speech acts in sport:. *Int. Rev. Sociol. Sport*.

Jane G. Stout and Nilanjana Dasgupta. 2011. When he doesn't mean you: Gender-Exclusive language as ostracism. *Pers. Soc. Psychol. Bull.*

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *ACL*.

Michael Warner. 1991. Introduction: Fear of a queer planet. *Social Text*, (29):3–17.

Monique Wittig. 1979. The straight mind. *The Future of Difference (discours liminaire en conférence universitaire)*, vol. 3, t. 3. New York, Barnard Center for Research on Women, coll.« Scholar and Feminist / VI.

Monique Wittig. 1980. The straight mind. *Feminist Issues*, 1(1):103–111.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral word embeddings.