HOMO-MEX: A Mexican Spanish Annotated Corpus for LGBT+phobia Detection on Twitter

Juan Vásquez[1] and **Scott Thomas Andersen** [2]
Posgrado en Ciencia e Ingeniería de la Computación
Universidad Nacional Autónoma de México
juanmv@comunidad.unam.mx and stasen@comunidad.unam.mx

Gemma Bel-Enguix [3] and
Sergio-Luis Ojeda-Trueba [5]
Instituto de Ingeniería
Universidad Nacional
Autónoma de México
gbele@iingen.unam.mx and
sojedat@iingen.unam.mx

Helena Gómez-Adorno [4]
Instituto de Investigaciones en
Matemáticas Aplicadas y en Sistemas
Universidad Nacional
Autónoma de México
helena.gomez@iimas.unam.mx

Abstract

In the past few years, the NLP[1] community has actively worked on detecting LGBT+Phobia in online spaces, using textual data publicly available Most of these are for the English language and its variants since it is the most studied language by the NLP community. Nevertheless, efforts towards creating corpora in other languages are active worldwide. Despite this, the Spanish language is an understudied language regarding digital LGBT+Phobia. The only corpus we found in the literature was for the Peninsular Spanish dialects, which use LGBT+phobic terms different than those in the Mexican dialect. For this reason, we present Homo-MEX, a novel corpus for detecting LGBT+Phobia in Mexican Spanish. In this paper, we describe our data-gathering and annotation process. Also, we present a classification benchmark using various traditional machine learning algorithms and two pre-trained deep learning models to showcase our corpus classification potential.

1 Introduction

LGBT+Phobia[2] is a global problem (Arimoro, 2022). Among the consequences faced by the LGBT+ community are substance abuse disorders among its members (Wallace and Santacruz, 2017), (Burkhalter, 2015), disproportionate mental health problems (Lozano-Verduzco et al., 2017) (MONGeLLi et al., 2019), discrimination in the labor markets (Quintana, 2009) (Ng and Rumens, 2017), denial of access to education and health services (Hatzenbuehler et al., 2017) (Ayhan et al., 2020), and lack of human rights (López, 2017) (Ungar, 2000) (Peck, 2022).

In recent years, NLP has greatly advanced its methods for detecting hate speech in online communities (Poletto et al., 2021).

Therefore, in order to detect LGBT+Phobia in social networks, specifically on Twitter, we created a corpus designed for this task. To the best of our knowledge, no other corpora focused on LGBT+Phobia in Mexican Spanish have been created so far. This can be very useful for NLP purposes because it is well-known that Mexican Spanish has a specific lexicon and pragmatics. Because of this, it would be valuable to have NLP systems specializing in this Spanish variant.

The corpus we present includes public tweets scraped using Twitter's API that includes keywords that we expect will be used in LGBT+phobic contexts. We gathered a list of nouns used to refer to the LGBT+ community. Then, we scraped nearly ten thousand tweets that contained any of these nouns from the past two years. Thereafter, four annotators annotated each tweet as LGBT+phobic, not LGBT+phobic, or not related to the LGBT+ community. Finally, another group of four annotators identified the fine-grained LGBT+phobic type.

The main contributions of our work are the following:

1. We create and manually annotate a corpus of

[1] Natural Language Processing

[2] Any and all references to the LGBT+ community or LGBT+Phobia includes all members of the LGBTQIA+ community, that is, all sexual and gender minorities that deviate from the traditional gender-binary or the traditional heterosexual relationship and the discrimination they face for their identity.

The 7th Workshop on Online Abuse and Harms (WOAH), pages 202–214
July 13, 2023 ©2023 Association for Computational Linguistics

tweets in Mexican Spanish based on a lexicon of LGBT+ terms[3].

2. We present various supervised classification models that could guide efforts towards the detection of online LGBT+Phobia; specifically, LGBT+Phobia in Mexican Spanish.

The rest of this paper is organized as follows. Section 2 surveys related literature and similar experiments. Section 3 describes the construction of our corpus. Section 4 details the methodology of the classification experiments. Section 5 discusses the results of the experiments. Finally, Section 6 describes experimental adjustments we would like to make in future experiments and closes the paper with conclusions. Appendix A provides a brief data statement to give insight into ethical considerations of the annotation process.

2 Related Work

Recent work explored using NLP to detect bullying, hate speech, violence, and aggressiveness. State-of-the-art models were developed for general-purpose hate detection and hateful content directed at a particular group. For these models to be developed, large quantities of data are essential. Recent data sets have emerged that seek to annotate hate towards marginalized groups, the majority being in English. Although multi-language models and data sets exist, having language-specific models and data (del Arco et al., 2021) is demonstrably helpful.

To the best of our knowledge, very few corpora have yet been developed to classify homophobic comments in Spanish in online communication. We seek to bridge this gap. First, we will describe relevant work in NLP, more specifically, models that are used for sentiment analysis, and identification of harassment and hate. Then, we present socially conscious work that seeks to be more inclusive and detect language discriminating against minority groups. Finally, we discuss works that include Spanish classification models.

2.1 NLP Models for general hate and abuse

Recent work in the NLP community seeks to detect harassment, bullying, and hate to improve the safety and quality of online spaces. In this section we present related work on sentiment analysis followed by hate detection.

2.1.1 Works for sentiment analysis

Models have been proposed to analyze sentiments in text for use in online platforms. For example, Demszky et al. (2020) includes a dataset of Reddit comments labeled with up to 27 emotions. Buechel et al. (2018) uses deep learning to learn emotion on data severely limited in size. They find that emotion can be successfully predicted even with models trained on very small data sets.

2.1.2 Works for hate detection

Plenty of work has come forth for the detection of hate speech and abusive language in Social Media (Lee et al., 2018; Kshirsagar et al., 2018; Jarquín-Vásquez et al., 2021).

Dinu et al. (2021) explores the use of pejorative language in Social Media, the context-dependent language used with a negative connotation. Similarly, discriminating language does not necessarily take the form of slurs but depends highly on the context of the comment.

Recent works like ElSherief et al. (2021) ElSherief et al. (2021) present a corpus of tweets as a benchmark for understand ing implicit rather than explicit hate speech.

Finally, HATECHECK (Röttger et al., 2021) provides functional tests for evaluating Hate speech detection models. These tests exposed key weaknesses and biases in state-of-the-art hate detection models.

2.2 Socially Conscious work in the NLP community

Socially conscious work has been made to detect racially, gender, or sexually inspired hate to make online spaces more inclusive. First, we will consider explicitly gender and racial bias, and following this, we will consider LGBT+-specific hate.

This is vital as Xu et al. (2021) demonstrates that standard detoxifying techniques can disproportionately affect generated text from minority communities. For example, by falsely flagging common identity mentions such as "gay" or "Muslim" because the model has learned to associate them with toxicity.

2.2.1 Hate and Bias

Hate and bias present in online spaces are harmful to minority and marginalized communities, but

[3] The link to the Github repository with the IDs and labels of the tweets will be available after the publication of this paper

recent efforts have been proposed to detect and address hateful and biased speech. Fraser et al. (2021) proposes the Stereotype Content Model in NLP adopted from social psychology to represent stereotypes along two axes, warmth, and competence. Their model takes words directed at a particular group and scores them on these dimensions. In addition, they discuss how to use this information to produce anti-stereotypes. Meanwhile, Sun and Peng (2021) creates an event-based dataset of gender bias from Wikipedia articles. They demonstrate that entries on females tend to include personal life events in career sections but not in the career sections for men. Meanwhile, more career-related achievements, such as awards, can be found in the personal life section for men but not for women. These subtle placements of events relevant to the person of interest are indicative of a gender bias in Wikipedia articles. Sheng et al. (2021) explores bias in Natural Language Generation tasks and provides a survey that explores how data and techniques can lead to bias in automatically generated text. They discuss how data, model architecture, methods for decoding, and even evaluation methods can produce a biased model.

An example of this bias, Excell and Moubayed (2021), demonstrates that using exclusively male annotators for a dataset of toxic comments yields weaker results than using exclusively female annotators. Combating this and keeping in mind that the scope of our work is LGBT+phobic content, we gathered annotators that identified as both male and female heterosexuals and members of the LGBT+ community. We also took special care to include annotators from various sexes so that each annotated subset of tweets with diverse representation of gender orientation and sexual identity (Section 3.2).

2.3 LGBT+ specific work

We wish to explore discrimination in natural language specific to the LGBT+ community. Several recent efforts have analyzed what kind of discrimination gender and sexual minorities face. For example, Gámez-Guadix and Incera (2021) addresses the sexual victimization of LGBT+ adolescents in online spaces, finding that many adolescents face gender and sexual-based victimization and receive unwanted sexual attention.

CH-Wang and Jurgens (2021) analyzes nearly 100 million tweets and Reddit comments to note the change of lexical variables indicative of support of gender and sexual minorities, finding that language use changes for community members who feel more accepted. They find that people shift from gender-neutral terms like "partner" to gender-specific terms like "husband" in places where marriage equality acts were enacted. Meanwhile, Khatua et al. (2019) analyzed tweets in India following the legalization of gay marriage. They found that tweets in support centered around justice and equality, while opposing tweets saw the decision as a threat to traditional Indian culture.

Hudhayri (2021) analyzes harassment toward Arab LGBTs in cyberspaces. They investigate semiotic harassment, which studies hidden connotations of harassment shared by language users.

Chakravarthi et al. (2021) generate a data set of multilingual transphobic and homophobic Youtube comments and use a diverse categorical labeling system to determine if the comment is homophobic or transphobic, specifying if it is derogatory or threatening, they even include labels for counterspeech and hope speech. Vargas et al. (2022) build a corpus of 7,000 Brazilian documents. Their corpus was annotated for a binary classification task (offensive versus non-offensive comments), and for a fine-grained classification task depending on the level of offensiveness found in the documents labeled as “offensive” (highly, moderately, and slightly offensive). Furthermore, the authors annotated the documents in nine classes, depending on the perpetrators of the hate speech found in their documents (xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology for the dictatorship, antisemitism, and fatphobia).

2.4 Hate Speech Identification in Spanish

On 2021, the PAN at CLEF Initiative organized the shared task *Profiling Hate Speech Spreaders on Twitter 2021*, which focused on identifying hate speech against people based on their race, color, ethnicity, gender, sexual orientation, nationality, religion, or another characteristic on Twitter (Bevendorff et al., 2021). The participants were given a dataset with tweets in English and Spanish and had to classify them into two classes. The highest accuracy obtained by the participants was 73.0% for tweets in English and 85.0% for tweets in Spanish (Rangel et al., 2021).

The IberLEF 2021 organized various shared tasks on *Harmful Information*. The first one, MeOffendEs@IBERLEF 2021, aimed at classifying offensive language and its categories in various Spanish dialects (Plaza-del Arco et al., 2021). Four subtasks were proposed in this shared task, all aimed at identifying *offensive language*. The organizers created a corpus made up of "multiple social networks and a diversity of variants of Spanish". The second shared task was EXIST (Rodríguez-Sánchez et al., 2021). Its goal was to identify online sexism. For this shared task, the participants were provided a dataset comprised of 6,977 tweets in English and Spanish. They had to perform two tasks: first, a binary classification of the tweets, then, a categorization of the type of sexism identified in the tweets. The highest accuracy obtained on the binary classification was 78.04%, while the classification among the types of sexism obtained an accuracy of 65.77%. The third shared task was DETOXIS (Taulé et al., 2021). This task aimed to "detect toxicity in comments posted in Spanish in response to different online news articles related to immigration". The highest F1 measure obtained in the first subtask, the toxicity detection task, aimed at performing a binary classification among the classes "toxic" and "non-toxic", was 85.16%. In contrast, the corresponding highest F1 for the second subtask, the toxicity level detection task, consisting of four labels, was 89.29%.

Similar efforts were the two shared tasks, Language Technology for Equality, Diversity, Inclusion (LT-EDI, ACL 2022); and SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. The organizers of the first shared task focused on the automatic classification of Youtube comments in English and Tamil, labeled as transphobic and homophobic (García-Díaz et al., 2022), while the latter aimed at classifying hateful speech in a binary classification task and identifying if the targets of the hateful messages were single individuals or groups of people (Basile et al., 2019).

3 Corpus for LGBT+Phobia Detection in Mexican Spanish

This section describes the methodology for collecting data, our annotation process, and the challenges we faced. We also report the agreement between annotators.

3.1 Data Collection

Using Twitter's API, we collected publicly posted Spanish tweets originating from Mexico. The Twitter API supposedly collects public tweets randomly, and we can expect the grand majority of these tweets will be in Mexican Spanish by native speakers. However, speakers of other backgrounds speaking other languages or variants of the language may appear.

We annotated a large set of tweets that contained any noun indicative of the LGBT+ community. These terms were collected by linguistic students tasked with finding every noun used in Mexican Spanish about the LGBT+ community. These terms were collected from social networks like Twitter, Facebook, Instagram, and TikTok to study social media discourse. We selected the most representative lexicon. Also, we contemplate the variations each term could have; particularly in the Mexican LGBT+ community, these nouns have appreciative inflections or inflections related to gender. For example, for gender, the noun *joto* could inflect in *jote*, *jotx*, and *jota*. In the appreciative case, we could derive forms such as *jotito*, *jotón*, *jotite*, etc.

The lemmas of the selected terms, along with their translation and frequency of appearance, can be found in Table 1, and the full Table can be found in the Github repository [4].

Having defined these search terms and variations, we scrape tweets using the Twitter API and filter them depending on their geolocation metadata.

The tweets were scraped between the dates 01-01-2012 to 01-10-2022 (day-month-year format) to ensure that we had a vast and diverse corpus of tweets. We obtained 706,886 unique tweets and annotated 11,000 from the ten year span, half of them from verified accounts – before the monetization of account verification on the platform – and the half randomly selected from the tweets published by unverified accounts.

3.2 Annotation Process

Having gathered and filtered tweets, we sought annotators to begin the annotation process. We collected annotators that were both heterosexual and members of the LGBT+ community to ensure a diverse set of perspectives were used when labeling the tweets. Before the annotation process, we had a group meeting with the annotators, and we discussed various example tweets and how they interpreted them. Then we launched a practice run and discussed the results together. We added a simple tutorial to the platform that gave some of these

[4] https://github.com/juanmvsa/HOMO-MEX

Keyword	Translation	Count	Keyword	Translation	Count	Keyword	Translation	Count
Bi	Bi(sexual)	3330	Trans	Trans	3245	Gay	Gay	1811
Loca	Crazy (fem.)	1116	Puto	Whore	1102	Homosexual	Homosexual	1049
Joto	Faggot	1008	Lesbiana	Lesbian	827	Drag	Drag queen	597
Marica	Faggot	537	Vestida	Dressed Up	458	Bisexual	Bisexual	336
Maricón	Faggot	487	Transexual	Transexual	290	Transformer	A Trans person	279
Transgénero	Transgenedered	227	Travesti	Transvestite	226	Queer	Queer	202
Lencha	Lesbian	181	Mayate	Lesbian	179	Puñal	Gay man	114
Rarx	Strange	108	No binario	Non-binary	79	Clostera	Closeted person	75
Afeminado	An Effeminate	73	Intersexual	Intersexual	58	Pansexual	Pansexual	54
Asexual	Asexual	43	Machorra	Lesbian	40	Cuir	Queer	28
Femboy	Femboy	19	Tortilla	Tortilla	11	Trapito	Little rag	7
Crossdresser	Crossdresser	7	Sáfica	Safic	6	Muxhe	Muxhe	4
Género fluido	Gender Fluid	4	Arcoiris	Rainbow	4	Demisexual	Demisexual	3
Enby	Non-Binary	3	Hombre Con Falda	Man with a Skirt	2	Transformista	Trans person	2
Tijeras	Scissors	2	Panes	Pan	2	Mariposon	Faggot	1
Lechugona	Lesbian	1	Bigénero	Bigender	1			

Table 1: The following table contains slurs against the LGBT+ community and may be offensive to some readers. The number of times each keyword, or their inflections, appear in the corpus. We list the search term, the English translation, and the number of tweets they appear in. Some terms were removed because they were too saturated with their non-LGBT+ interpretation: bicicleta (bicycle), and tortilla. Tortilla, however, still appears in tweets that contain another search term.

examples and the labels the group created and clarified questions many annotators had. All annotators in this meeting had to go through this tutorial to refresh their memory before beginning the annotation. At some points, we had to add additional annotators to replace some who dropped out, and we required them to go through this tutorial as well and asked them to reach out with any questions or concerns; this is to ensure consistent understanding of the annotation process for those who could not attend the initial meetings. The tutorial also formally defined some terms such as *LGBT+phobia* and *Transphobia* and included some questions that they were required to answer to proceed to ensure that they were paying attention to the content. We include more information on the annotators in the Data Statement.

3.3 Annotation Schema

Here we explain the methodology for labeling the tweets and how we measured agreement between the annotators.

The annotators labeled the 11,000 tweets as "LGBT+phobic", "Not LGBT+phobic", and "irrelevant to the LGBT+ community". In this task, the annotators could only select one category. All tweets labeled as "LGBT+phobic" were later passed through an additional annotation process that identified the type of LGBT+phobia. In the second stage, the labels were "gayphobia", "lesbophobia", "biphobia", "transphobia", and "other lgbt+phobic content". Although *gay* is an umbrella term that encompasses much of the LGBT+ community, for the purposes of this annotation, we requested that the annotators only use this label if the tweet contained LGBT+phobic content towards homosexual cis-males to best contrast with the other labels. In this task, the annotators were allowed to annotate the tweets with all labels that applied because one tweet could have LGBT+phobic content towards multiple groups.

In the LGBT+phobia detection task, we requested that if a tweet could be seen as LGBT+phobic if the author does not belong to the community and not LGBT+phobic if the author is LGBT, the annotators give the benefit of the doubt to the author. Therefore, the dataset did not overuse the LGBT+phobic label when much of the discourse within the community can be seen as ironically LGBT+phobic without true intent of harm towards the LGBT+ community.

The annotators used a custom annotation platform that presented the tweets to them in random order and ensured that their responses were anonymized while verifying that each tweet is labeled by four annotators, two members of the LGBT community and two heterosexual, male, and female.

In the LGBT+phobia identification set, a label was selected if it had the majority of the votes. All tweets tied were presented to a different set of annotators to be re-annotated. Any tweets still presented a tie after this were assigned a final label based on a final specialized annotator's decision.

In the type of LGBT+phobia identification set, any label that had at least half of the annotators' votes was selected as a label for the tweet. In this task, the tweet can have multiple labels, such as "gayphobia", "lesbophobia" and "transphobia".

3.4 Annotation Results

After the annotation was completed, we examined the agreement of the annotators for each subset of the corpus, using Fleiss' Kappa. This information is available in Table 2.

For the detection subset we see a moderate agreement among all groups in the phobia detection task, and in the re-annotated tweets that had tied. We calculate the agreement among LGBT+ and Non-LGBT+ annotators, and compare it to the agreement among those of female or male sex, as well as among all annotators.

The fine-grained annotation agreements are not as consistent. We see that there is much more agreement among Non-LGBT+ annotators and Male annotators in every category of LGBT+phobia. LGBT+ annotators and Female annotators show the most disagreement in the annotation of gayphobia and Other types of LGBT+phobia. We hypothesize that this could be from inconsistent interpretations of language use in LGBT+ sub-communities that male and non-LGBT+ annotators may be less exposed to, keeping in mind that a group being in agreement does not necessarily mean they are correct.

3.4.1 Examples

With the tweets annotated, here we will provide a few examples of tweets and their labeling and a rough translation. Warning: these tweets could include distressing language and slurs against the LGBT+ community that may harm some readers.

LGBT+phobic Tweets Here are two examples of tweets that were labeled as LGBT+phobic. *"De que me sirve tener amigos gays si no me sirven para consejos de moda #badgayfriends"*, roughly translated to *"It is usless have gays friends if they dont give me fashion advice #badgayfriends"*. The author of this tweets assume that all the homosexuals know about fashion, a frequent stereotype that is also present in the hashtag. Another example is *"Lo siento, soy muy marica para el dolor)':"*, translated again roughly as *"Im sorry, Im such a fag when it comes to pain)':"*. Here the author relates weakness with the LGBT+ community.

Non LGBT+phobic Tweets Here we will include a few examples of tweets that were labeled as not having LGBT+phobic intent.*"Estados Unidos levanta la prohibición para que homosexuales donen sangre"*, translated to *"The United States lifts ban on homosexuals donating blood"*. Another example is *"Entonces lo que anda(mos) haciendo las viejas trans es crearnos mujeres COMO SE LE ENSEÑA AL NIÑO que es una mujer (objeto, sexuada, sumisa)"*, translated again roughly as *'So what we old trans are doing, We are making ourselves women as HOW BOYS ARE TAUGHT that a woman is (objectified, sexualized, submissive)'"*. Here the author employs *trans* to refer to themselves naturally.

Tweets with low agreement The following tweets had low agreement in the detection task. *"Ah verga es un duende? Yo pensaba era un alíen asexual"*, or in English, *"Ah fuck they're an elf? I thought they were an asexual alien."*, this tweet was labeled as LGBT+phobic. *"No, Sifo, no. O sea, no mames. No soy una puta. Qué te pasa. Si quieres que te la chupe, me vas a tener que pagar."*, which translates as *"No, Sifo, no. I mean, quit fucking with me. I'm not a whore. If you want me to suck it, you'll have to pay."* which was finally labeled as not relevant to the LGBT+ community.

3.5 Challenges to Annotation

One challenge we faced during the creation of Homo-MEX was the annotation process. Even though we had various annotators that were members of the LGBT+ community and/or were very aware of the issues faced by the LGBT+ Mexican community, the annotator inter-agreement was not very high. We attribute these results to the difficulty of differentiating between irony, resignification, appropriation of slurs, and humor inside the LGBT+ community, especially when the context may not be available. This limitation is important, however, because it best aligns with the circumstances of automatic LGBT+phobia detection based on just the tweets' textual content.

Another potential limitation could be the difficulty in counterbalancing the internalized stereotypes that the annotators might have. This has proven to influence the annotation behaviors (Davani et al., 2023).

The annotator agreement is especially low for the label "Other" in the fine-grained classification task. We suppose that the label may not be well

Detection Subset	Kappa				
	LGBT+	Non-LGBT+	Male	Female	All
Phobia Detection	0.449	0.371	0.392	0.474	0.430
Tie Break	0.517	0.369	0.416	0.409	0.465

Fine Grained Subset	Kappa				
	LGBT+	Non-LGBT+	Male	Female	All
Gayphobia	-0.055	0.732	0.789	-0.087	0.316
Lesbophobia	0.691	0.656	0.723	0.572	0.665
Biphobia	0.205	0.565	0.495	0.315	0.419
Transphobia	0.650	0.743	0.779	0.638	0.700
Other	-0.306	0.353	0.422	-0.322	-0.027

Table 2: We employ Fleiss' kappa to analyze the agreement among the annotators. More information can be found in Section 3.3. The group *Phobia Detection* refers to the annotation task identifying tweets that did or did not contain LGBT+phobia or were irrelevant to LGBT+ discourse. The group *Tie Break* is the agreement among the annotators who reclassified the tweets that tied in labels from the previous group. Finally, the second table represents the agreement for each LGBT+phobia category in the Fine Grained data set.

defined, or more nuanced types of LGBT+phobia may not be as easy to identify.

4 Experiments on the HOMO-MEX Corpus for LGBT+phobia Detection

To evaluate the performance of various classifiers on our corpus, we performed several experiments using two main approaches: traditional machine learning methods and deep learning architectures. We describe these experiments in this section.

The HOMO-MEX corpus consists of two overlapped subsets. The first subset is comprised of those tweets that can be either "LGBT+Phobic" (LP), "Not LGBT+Phobic" (NLP), and "irrelevant to the LGBT+ community" (I). On the other hand, the second subset contains the LGBT+Phobic tweets that were multi-labeled as "Lesbophobic" (L), "Gayphobic" (G), "Biphobic" (B), "Transphobic" (T), and "Other" (O). For conciseness, we will refer to the first subset as "LGBT+Phobia detection", and the second as "fine-grained classification". Both LGBT+Phobia detection and fine-grained classification subsets were split into train and test partitions. The resulting size and distribution of labels in each partition are shown in Tables 3 and 4. In table 4, the total of the train and test partitions is equal to 862 and 477, respectively, even though the addition of the tweets with every label (L, G, B, T, O) does not add to the counts since the tweets in this partition can have more than one label at a time. This allows the number of labels to be greater than the total size of the train and test partitions.

Partition	LP	NLP	I	Total
Train	862	4,360	1,778	7,000
Test	477	2,493	1,030	4,000
Total	1,339	6,853	2,808	11,000

Table 3: Size and label distribution for the LGBT+Phobia detection subset.

Partition	L	G	B	T	O	Total
Train	72	714	10	79	64	862
Test	34	414	3	38	32	477
Total	106	1,128	13	117	96	X

Table 4: Size and label distribution for the fine-grained classification subset.

4.1 Traditional Machine Learning Approach

Initially, we performed several pre-processing steps to the corpus. The first step in this process was the removal of stopwords using nltk's lexicon[5]. Then, we removed all diacritic characters, digits, and all other characters that were not a letter, or an underscore. Following, we tokenized the tweets using spaCy's small Spanish model, $es\_news\_core\_sm$[6]. Finally, we generated the features for the different machine-learning algorithms. To achieve this, we made use of the bag-of-words algorithm and TF-IDF weighting scheme as implemented in scikit-learn (version 0.23.2) [7] .

[5] https://github.com/xiamx/node-nltk-stopwords/blob/master/data/stopwords/spanish

[6] https://spacy.io/models/es

[7] https://scikit-learn.org/stable

4.2 Pre-trained Deep Learning Models Approach

Using both subsets (LGBT+Phobia detection and fine-grained classification), we fine-tuned various pre-trained large language models for classification. No pre-processing steps were performed in these experiments. The large language models that we used for these classification experiments were `bert-base-multilingual-cased` (Devlin et al., 2018), `bert-base-multilingual-uncased` (Devlin et al., 2018), `beto-cased` (Cañete et al., 2020), and `beto-uncased` (Cañete et al., 2020). We used hugging face's `transformers` (Wolf et al., 2019) library for their implementation[8].

5 Results and Discussion

We performed classification experiments using Naive Bayes, SVM, Logistic Regression, and Random Forest classifiers. Table 5 shows the metrics obtained using the LGBT+Phobia subset, and Table 6 shows the classification metrics obtained using the fine-grained subset. In addition, we used four BERT models to classify the tweets in both the LGBT+Phobia detection and fine-grained classification subsets. The results of these experiments can be observed in Table 7 for the LGBT+Phobia detection subset and in Table 8 for the fine-grained classification subset. We follow the PT1 method explained in Tsoumakas and Katakis (2007) to evaluate the fine-grained classification models. The PT1 method consists of splitting a classification problem (with $L = [A, B, C, D, E]$ labels) into a classification problem with $M = L \bigcup N$ labels, where $N = [\neg A, \neg B, \neg C, \neg D, \neg E]$. Then, the classification is treated as five binary subtasks, one for each label and its negation. For example, the first binary classification subtask would be with the labels $[A, \neg A]$, the second binary classification with the labels $[B, \neg B]$, and so on. Once the five metrics, one for each subset of labels, were generated, the average between them was computed. Those averages are reported in Tables 6 and 8.

Among the classical machine learning algorithms, SVM performs the best among almost all metrics in both partitions. Beto-cased produces the highest classification metrics in the LGBT+Phobia detection subset, while bert-base-multilingual-uncased outperforms the other bert-based models in the fine-grained classification subset. These results demonstrate that more work must be done on automatically classifying LGBT+phobic speech in Mexican Spanish.

6 Conclusion and future work

Detecting LGBT+Phobia using current NLP techniques is still an open task with much work left to do. The paper's contribution is twofold: first, we elaborate on a resource to study the topic in Mexican Spanish. Additionally, we test traditional ML methods, as well as BERT-based techniques, to identify LGBT+Phobia.

The corpus has been designed by filtering tweets with specific keywords related to the LGBT+ community in Mexico. Such tweets contain many references to LGBT+Phobia. However, surprisingly, there is more hateful speech when referring to the masculine gay community. Looking at the tweets with feminine terms, we see that many were written by women inside the community. This implies a different problem, the general invisibility of women, that should be tackled in the more general framework of sexism.

In the future, we hope to continue to expand the dataset to include more tweets with even more diverse terms to represent all members of the LGBT+ community. At present, many of the tweets marked as discriminatory only exhibit homophobia towards men.

A future dataset should include a more profound labeling procedure that can reduce ambiguity for the annotators and provide more information using a non-binary labeling system. Future approaches can include the categories of derogatory, threatening, humor remark and apparently neutral comment, among others.

Future papers should create a more representative dataset of Mexican Spanish tweets with a more thorough labeling system. Moreover, it will be interesting to the collection corpora in several variants of Spanish. With this, we plan to start a dialectal approach to the problem.

Furthermore, for automatic classification tasks, NLP practitioners should consider including lexicon-informed approaches for the generation of context-aware features for their classifiers, since this has proven its effectiveness in the case of hate speech detection from Brazil(Vargas et al., 2021). Finally, we wish to reiterate that further computational efforts against hate speech should always

[8] https://huggingface.co/docs/transformers/v4.28.1/en/model_doc/bert#transformers.BertForTokenClassification

Classification algorithm	Accuracy	Precision	Recall	F1-score
Naive Bayes	0.6885	0.8542	0.4244	0.4127
SVM	**0.8452**	0.7955	**0.7519**	**0.7670**
Logistic regression	0.8447	**0.8274**	0.7244	0.7592
Random forest	0.8302	0.7965	0. 7037	0.7349

Table 5: Classification results experiments using traditional ML algorithms on the LGBT+Phobia detection subset.

Classification algorithm	Accuracy	Precision	Recall	F1-score
Naive Bayes	0.9287	0.9643	0.5000	0.4813
SVM	**0.9589**	**0.9700**	**0.6558**	**0.6909**
Logistic regression	0.9312	0.9156	0.5122	0.5048
Random forest	0.9534	0.9648	0.6281	0.6622

Table 6: Classification results experiments on the fine-grained classification subset.

Classification algorithm	Accuracy	Precision	Recall	F1-score
bert-base-multilingual-uncased	0.8577	0.8558	0.8577	0.8566
bert-base-multilingual-cased	0.8492	0.8488	0.8485	0.8494
beto-cased	**0.8600**	**0.8592**	**0.8589**	**0.8600**
beto-uncased	0.8552	0.8554	0.8555	0.8552

Table 7: Classification results using BERT models on the LGBT+Phobia detection subset.

Classification algorithm	Accuracy	Precision	Recall	F1-score
bert-base-multilingual-uncased	**0.7815**	**0.9354**	**0.7815**	0.7396
bert-base-multilingual-cased	0.7614	0.8417	0.7614	**0.7422**
beto-uncased	0.7765	0.7713	0.7765	0.7403
beto-cased	0.7710	0.7879	0.7711	0.7416

Table 8: Classification results using BERT models on the fine-grained classification subset.

take into account LGBT people's experiences while designing their experiments. This, in recognition that hateful discourses against this population are often constructed by intersecting power structures –such as the symbolic discourses that produce the "immoral, defective, and inferior LGBT Individual" – which further limit the collaboration between the LGBT+ population and Academia in the battle against hate speech.

References

Augustine Edobor Arimoro. 2022. *Global Perspectives on the LGBT Community and Non-Discrimination*. IGI Global.

Cemile Hurrem Balik Ayhan, Hülya Bilgin, Ozgu Tekin Uluman, Ozge Sukut, Sevil Yilmaz, and Sevim Buzlu. 2020. A systematic review of the discrimination against sexual and gender minority in health care settings. *International Journal of Health Services*, 50(1):44–61.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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.

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.

Janek Bevendorff, Berta Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, et al. 2021. Overview of pan 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 419–431. Springer.

Sven Buechel, João Sedoc, H. Andrew Schwartz, and Lyle H. Ungar. 2018. Learning neural emotion analysis from 100 observations: The surprising effectiveness of pre-trained word representations. *CoRR*, abs/1810.10949.

Jack E Burkhalter. 2015. Smoking in the lgbt community. In *Cancer and the LGBT Community*, pages 63–80. Springer.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020(2020):1–10.

Sky CH-Wang and David Jurgens. 2021. Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender. *CoRR*, abs/2109.11061.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, and John Phillip McCrae. 2021. Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments. *Natural Language Engineering*, page 44.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Flor Miriam Plaza del Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. Number: arXiv:2005.00547 arXiv:2005.00547 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. A Computational Exploration of Pejorative Language in Social Media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *CoRR*, abs/2109.05322.

Elizabeth Excell and Noura Al Moubayed. 2021. Towards equal gender representation in the annotations of toxic language detection.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model.

José García-Díaz, Camilo Caparros-Laiz, and Rafael Valencia-García. 2022. UMUTeam@LT-EDI-ACL2022: Detecting homophobic and transphobic comments in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 140–144, Dublin, Ireland. Association for Computational Linguistics.

Manuel Gámez-Guadix and Daniel Incera. 2021. Homophobia is online: Sexual victimization and risks on the internet and mental health among bisexual, homosexual, pansexual, asexual, and queer adolescents. *Computers in Human Behavior*, 119:106728.

Mark L Hatzenbuehler, Andrew R Flores, and Gary J Gates. 2017. Social attitudes regarding same-sex marriage and lgbt health disparities: Results from a national probability sample. *Journal of Social Issues*, 73(3):508–528.

Khalid Hudhayri. 2021. Linguistic harassment against arab lgbts on cyberspace. *International Journal of English Linguistics*, 11(4).

Horacio Jarquín-Vásquez, Hugo Jair Escalante, and Manuel Montes. 2021. Self-contextualized attention for abusive language identification. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 103–112, Online. Association for Computational Linguistics.

Aparup Khatua, Erik Cambria, Kuntal Ghosh, Nabendu Chaki, and Apalak Khatua. 2019. Tweeting in support of lgbt? a deep learning approach. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '19, page 342–345, New York, NY, USA. Association for Computing Machinery.

Rohan Kshirsagar, Tyus Cukuvac, Kathleen R. McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. *CoRR*, abs/1809.10644.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative Studies of Detecting Abusive Language on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, Brussels, Belgium. Association for Computational Linguistics.

Jairo Antonio López. 2017. Los derechos lgbt en méxico: Acción colectiva a nivel subnacional. *European Review of Latin American and Caribbean Studies/Revista Europea de Estudios Latinoamericanos y del Caribe*, 104:69–88.

Ignacio Lozano-Verduzco, Julián Alfredo Fernández-Niño, and Ricardo Baruch-Domínguez. 2017. Asociación de la homofobia internalizada con indicadores de salud mental en personas lgbt de la ciudad de méxico. *Salud mental*, 40(5):219–226.

Francesca MONGeLLi, Daniela Perrone, Jessica BaLDUcci, Andrea Sacchetti, Silvia Ferrari, Giorgio Mattei, and Gian M Galeazzi. 2019. Minority stress and mental health among lgbt populations: An update on the evidence. *Minerva Psichiatrica*.

Eddy Ng and Nick Rumens. 2017. Diversity and inclusion for lgbt workers: Current issues and new horizons for research. *Canadian Journal of Administrative Sciences*, 34(2):109–120.

Steven Peck. 2022. The criminal justice system and the lgbtq community: An anti-queer regime. *Themis: Research Journal of Justice Studies and Forensic Science*, 10(1):5.

Flor Miriam Plaza-del Arco, Marco Casavantes, Hugo Jair Escalante, M Teresa Martín-Valdivia, Arturo Montejo-Ráez, Manuel Montes, Horacio Jarquín-Vásquez, Luis Villaseñor-Pineda, et al. 2021. Overview of meoffendes at iberlef 2021: Offensive language detection in spanish variants. *Procesamiento del Lenguaje Natural*, 67:183–194.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Nico Sifra Quintana. 2009. Poverty in the lgbt community. *American Progress*.

Francisco Rangel, Gretel Liz De la Peña Sarracén, BERTa Chulvi, Elisabetta Fersini, and Paolo Rosso. 2021. Profiling hate speech spreaders on twitter task at pan 2021. In *CLEF (Working Notes)*, pages 1772–1789.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on wikipedia.

Mariona Taulé, Alejandro Ariza, Montserrat Nofre, Enrique Amigó, and Paolo Rosso. 2021. Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish. *Procesamiento del Lenguaje Natural*, 67:209–221.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Mark Ungar. 2000. State violence and lesbian, gay, bisexual and transgender (lgbt) rights. *New Political Science*, 22(1):61–75.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.

Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447.

Barbara C Wallace and Erik Santacruz. 2017. Addictions and substance abuse in the lgbt community: New approaches. *LGBT psychology and mental health: Emerging research and advances*, pages 153–175.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

A Appendix

Data Statement We follow the guidelines specified by (Bender and Friedman, 2018) for creating a Data Statement, which serves to help 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 on various topics in a digital medium. We use specific search terms that are common nouns to refer the LGBT+ community to help identify hateful speech against the community.

B. Language variety We scrape a set of tweets that contained desired keywords and were in Spanish with the specified region of Mexico to get language of this region. Also, we took in consideration possible inflection of the terms. Since all the data is collected from social media, this means that there could be present hashtags, mentions, gifs, videos, images, and emojis within the tweets, however only the text of the tweet was utilized for annotation.

C. Tweet author demographic The demographics of the authors is not available to us since we compiled the data using Twitter's data collection API. However, due to our sampling methods, we expect the tweets to come from the diverse set of authors of various ages, genders, nationalities, races, ethnicities, native languages, socioeconomic classes and education backgrounds that are to be expected to be found within Mexico.

D. Annotator demographic We selected annotators that self identified as members of the LGBT+ community and non-members. The demographic information is shown in Table 9.

Categories	**Data**
Age	22-35 years
Gender Identity	1 non-binary 6 women 5 men
Sex	6 female 6 male
Sexual Orientation	6 LGBT+ 6 Cis-Heterosexual
Native Languague	Spanish
Nationality	11 Mexican 1 Colombian
Residence	México City
Education level	University

Table 9: Annotator demographic

E. Speech Situation Each tweet may be on a different topic. Most of them are related to trends, 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.

H. Ethical Statements All tweets were uploaded only by their ID. The textual content was omitted to assure the privacy of the author and the username of the people that could be mention on the tweet. All scraped tweets were posted publicly and can be collected for academic use according to Twitter's privacy policy.

Also, all the annotators were informed about the task and what type of profile we pursued for the project. In the annotation guidelines, we warned the annotators that the tweets could be offensive and that they could leave the study at any time.