A Cross-Lingual Study of Homotransphobia on Twitter

Davide Locatelli
Technical University of Catalonia
Barcelona, Spain
davide.locatelli@upc.edu

Greta Damo
Bocconi University
Milan, Italy
greta.damo@studbocconi.it

Debora Nozza
Bocconi University
Milan, Italy
debora.nozza@unibocconi.it

Abstract

We present a cross-lingual study of homotransphobia on Twitter, examining the prevalence and forms of homotransphobic content in tweets related to LGBT issues in seven languages. Our findings reveal that homotransphobia is a global problem that takes on distinct cultural expressions, influenced by factors such as misinformation, cultural prejudices, and religious beliefs. To aid the detection of hate speech, we also devise a taxonomy that classifies public discourse around LGBT issues. By contributing to the growing body of research on online hate speech, our study provides valuable insights for creating effective strategies to combat homotransphobia on social media.

Warning: *this paper contains examples of offensive language.*[1]

1 Introduction

Despite significant advancements in laws and societal attitudes surrounding LGBT rights around the world, homotransphobia, which refers to the hatred and discrimination towards individuals who identify as lesbian, gay, bisexual, or transgender, remains a pervasive phenomenon across diverse cultures (Pousher and Kent, 2020). The prevalence and visibility of hate speech toward LGBT individuals have escalated in the age of social media, further exacerbating the challenge of combating such discriminatory behavior. Recent surveys reveal that a substantial proportion of LGBT individuals have fallen prey to online attacks through homotransphobic messages, posing a serious threat to their well-being.[2,3]

The fight against online homotransphobic speech can be aided by natural language processing (NLP) techniques. Automatic hate speech detection systems, in particular, have the potential to reduce the spread of harmful language flagging such content for removal. However, the task of detecting homotransphobic speech is far from simple, given the multifaceted nature of this phenomenon. In order to accurately identify it, detection methods must take into account cross-lingual factors and recognize the subtle nuances in how this form of intolerance manifests itself in different cultures.

Despite its social relevance and harmful effects, this phenomenon has received little attention from NLP researchers compared to other types of hate speech, such as aggression (Kumar et al., 2018), misogyny (Fersini et al., 2018, 2020, 2022), and racism (Waseem and Hovy, 2016; Lee et al., 2022). One of the main challenges for developing effective homotransphobic detection models is the scarcity of annotated data in this domain (Chakravarthi et al., 2021; Carvalho et al., 2022; Nozza, 2022) and the negative bias of NLP models regarding LGBT individuals (Nozza et al., 2022).

In this paper, we conduct a cross-lingual study to investigate public discourse surrounding LGBT issues on Twitter, to identify areas where homotransphobic speech persists. To achieve this, we analyze a vast corpus of tweets in seven languages using topic modeling and sentiment analysis. These techniques have been extensively used in observational studies (Dahal et al. 2019; Xue et al. 2020; Lyu et al. 2021, inter alia). We aim to offer a nuanced understanding of the emergence of different themes of homotransphobic speech across different languages. Additionally, we propose a taxonomy for categorizing this discourse, estab-

[1] Obfuscation was done with PrOf (Nozza and Hovy, 2022)
[2] https://www.glaad.org/smsi
[3] https://www.ustranssurvey.org/reports

Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 16–24

L	TOTAL	SAMPLE	POS	NEU	NEG
DE	44,889	25,000	15%	33%	52%
EN	1,070,280	25,000	32%	32%	36%
ES	164,451	25,000	11%	27%	62%
FR	93,395	25,000	18%	11%	71%
IT	59,830	25,000	22%	28%	50%
NO	5,036	5,036	15%	30%	54%
PT	38,070	25,000	12%	18%	71%

Table 1: Overview of the data by language (**L**). We report the number of tweets collected (**TOTAL**), the number of tweets used for analysis (**SAMPLE**), and the proportions of positive, neutral, and negative sentiment tweets with respect to the sample.

lishing a foundation for the development of more effective homotransphobic speech detection models. We maintain the project repository at https://github.com/MilaNLProc/crosslingual-analysis-homotransphobia.

2 Data

We examined seven languages – German, English, French, Italian, Norwegian, Spanish, and Portuguese – and collected tweets containing LGBT keywords. These included both neutral terms (e.g., "gay") and derogatory slurs (e.g., "f*ggot").

To ensure that our list of keywords is comprehensive and representative of the different linguistic contexts, we recruited native speakers for each language in our study. Moreover, we selected individuals who are familiar with the LGBT community and its terminology. Where possible, we included multiple native speakers per language from diverse backgrounds and regions.

Using Twitter's historical API, we retrieved around 1.5 million tweets from May to September 2022, which coincided with Pride Month celebrations that we expected to increase discussions on LGBT issues. We sampled 25,000 tweets for each language, except for Norwegian, which had fewer tweets. To ensure that our collection reflects a realistic distribution, we compared it with an estimate of the total number of tweets posted for each language in a week during the same period. The number of tweets for each language is summarized in Table 1. For more information on our keyword selection, preprocessing and methodology for estimating the number of tweets per week, refer to Appendix A.

3 Methodology

We extracted 10 topics for each language, using Contextualized Topic Modeling (CTM) (Bianchi et al., 2021). We then developed a taxonomy to characterize LGBT public discourse, consisting of five broad categories and several subcategories, described in Table 2. We used this to label topics with a unified framework. Two in-house annotators labeled each topic based on the top words and a sample of 100 tweets for each topic, translated in English using an automatic translation software [4]. The annotators resolved discrepancies through discussion.

To devise this taxonomy we employed a multi-round process of annotation. First, we conducted a review of relevant literature from social science studies to identify common themes (Bianchi 2014; Slaatten et al. 2015; la Roi and Mandemakers 2018; Johannessen 2021; Hartmann-Tews et al. 2021; Biancalani et al. 2022, inter alia). Next, we collected personal accounts from LGBT individuals, with a particular focus on their perception of LGBT public discourse. Based on these findings, we created an initial draft of the taxonomy that grouped the themes into categories. To ensure that the framework was as accurate as possible, the annotators used it to devise initial labels for the topics emerged from CTM. In cases where inconsistencies were found, we refined the taxonomy further, breaking down each category into subcategories. Tweets that were discovered to touch on subjects unrelated to LGBT issues were grouped into a distinct category named "Other / Irrelevant". For instance, tweets that were selected using a keyword with multiple meanings, some of which were not related to the LGBT community, were placed in this category.

We then used a pre-trained multilingual sentiment analysis classifier (Barbieri et al., 2022) to analyze the attitudes expressed in the tweets. Here, we employ sentiment as a soft proxy for homotransphobia, because no multilingual detection models have been developed to date and cross-lingual hate speech detection methods does not transfer across different targets and languages (Nozza, 2021). It is important to note that the sentiment of a tweet is not a perfect measure for identifying hate speech, since it can potentially capture other phenomena, overlook some forms of hate speech, and misinterpret benign language as hateful due to contextual nuances and subtleties of natural language. However,

[4] https://www.deepl.com/translator

CATEGORY	SUBCATEGORY	TOPICS	EXAMPLE
Gender and Sexuality	Gender roles and sexual identity	Societal expectations on gender / sex	*Trans women are not women*
	Language and terminology	Meaning of LGBT words	*You can't say f*ggot*
	Pornography	Pornographic content	*Click to see this s*ssy*
Prejudice	Cultural stereotypes	Homotransphobic beliefs	*Gays will burn in hell*
	Slurs and stigmatization	Insults using anti-LGBT words	*You're such a f*ggot*
Sociopolitical influences	Politics and policy	LGBT rights	*F*ck the Equality Act*
	Events and organizations	Promoting LGBT visibility	*Can't wait for Pride!*
	Legal issues	Legal challenges / advocacy efforts	*Sign this petition for gay rights...*
Cultural representation	Representation in media	LGBT portrayal in media	*The main character is gay*
	Anti-LGBT language in sports	Homotransphobic slurs in sports	*Your team plays like f*ggots*
Other / Irrelevant		Topics irrelevant to LGBT issues	*I smoked a f*g yesterday*

Table 2: A taxonomy to categorize public discourse on LGBT issues, organized into five categories, and several subcategories. **TOPICS** indicates the content of the discussions belonging to each category, along with an example.

SUBCATEGORY	DE	EN	ES	FR	IT	NO	PT
Gender roles and sexual identity	18	–	13	–	7	13	8
Language and terminology	29	12	4	17	10	26	13
Pornography	13	35	–	14	–	–	–
Cultural stereotypes	–	–	–	9	8	–	13
Slurs and stigmatization	13	18	21	–	16	–	20
Politics and policy	6	12	6	22	–	34	17
Events and organizations	–	–	–	–	39	19	–
Legal issues	21	–	24	–	–	8	13
Representation in media	–	–	26	–	9	–	–
Anti-LGBT language in sports	–	–	–	5	11	–	15
Other / Irrelevant	–	23	6	31	–	–	–

Table 3: Proportion (%) of tweets by subcategory and language, and corresponding sentiment. Values in the cells represent the percentage of tweets that fall into a particular subcategory (row) for a given language (column). When a category has no tweets, we denote this by –. The color coding indicates the primary sentiment of the tweets: red for negative, yellow for neutral, green for positive. The intensity corresponds to the proportion of tweets in that sentiment.

we still opted to utilize it as it can offer valuable insight into the distribution and frequency of hate speech, and provide a starting point for further investigation. The sentiment distribution for each language can be found in Table 1.

4 Results

In this section we describe the main findings by category, which are summarised in Table 3.

4.1 Gender and sexuality

Gender and sexuality are topics that vary widely across languages.

Gender roles and sexual identity Transgender issues are a common theme in German, Norwegian, and Spanish, as indicated by words such as "women" and "trans". However, these languages differ in the perspectives expressed. German and Norwegian focus on transgender women's experiences, while Spanish shows dismissiveness toward transgender identity, painting it as a way for men to avoid responsibility for sexual violence against women, leading to a more negative sentiment (66%) compared to German (57%) and Norwegian (51%).

German and Norwegian tweets also examine the social construction of gender roles with words like "men", "gender", "manliness". They also explore the intersectionality between LGBT and disabled communities with words like "disabled" and "diversity". Moreover, they discuss self-identification versus external labeling with words like "queer", "lesbian", "love". Spanish tweets touch on similar topics but less frequently, with fewer related words.

Language and terminology Transgender-related terminology is widely discussed on Norwegian Twitter. Most tweets (65%) express neutral or positive sentiments, and contain respectful and productive engagement with debates surrounding the appropriateness of trans-related words, such as "transsexual" versus "transgender". German and French Twitter discussions focus on broader LGBT terminology. German tweets often debate how to refer to LGBT individuals, including reclaiming terms like "f*g" or "gay". Despite a high negative sentiment (67%), this may reflect the discussed words rather than negative attitudes. French tweets frequently use irony and provocation when discussing LGBT language and definitions, along with slurs and offensive language. Consequently, 80% of these tweets have a negative sentiment.

Pornography Pornography is prevalent in English, German, and French but not in other languages. These tweets typically include descriptions, links to content, and hashtags with explicit language. The English language global dominance may account for its high volume of pornographic tweets. Sentiment analysis shows that most English and German tweets are neutral or positive (over 80% and 70%, respectively), while French ones are less so (51%). This may not be accurate due to the sentiment analysis model not being well trained for pornographic tweets.

4.2 Prejudice

Prejudice and discrimination topics appear in all languages except Norwegian.

Cultural stereotypes Cultural stereotypes elicit negative sentiment in Portuguese, French, and Italian. Portuguese tweets mainly criticize the church's homophobia, with a highly negative sentiment. French and Italian tweets are classified as less negative, but they express more homophobic views, linking homosexuality to monkeypox, and opposing homosexual families.

Slurs and discrimination Homotransphobic slurs pervade tweets in all languages, except Norwegian. LGBT and non-LGBT individuals are equally targeted. Sex-related slurs are more prominent in English and German tweets, sometimes reclaimed by German LGBT people. English tweets also contain more pornography and less negativity (43%) than other languages (65-80%).

4.3 Sociopolitical influences

All languages contain tweets about social and political influence, especially Norwegian.

Politics and policy Politics and policy appears in all languages but Italian. French and Portuguese use homophobic slurs to attack right-wing politicians, with negative sentiment (87% and 74%). German tweets mock the idea that vaccines can lead a person to become gay, showing an interesting link to misinformation campaigns. English and Norwegian discuss legal rights for LGBT people, with neutral sentiment. Spanish tweets debate abortion rights and the deviance stigma of being gay.

Events and organizations Italian and Norwegian tweets mention LGBT events, mostly Italian (39%). This subcategory has mixed sentiment. In Italian, positive tweets use inclusive gender-neutral language, while negative ones lament the users' inability to join Pride parades for various reasons. Both Italian and Norwegian worry about LGBT safety after the Oslo shooting against Pride, pointing out that younger LGBT people are especially vulnerable. The dominant sentiment is negative, but mild (36% for both languages).

Legal Issues This category appears in German, Norwegian, Portuguese, and Spanish. All languages demand legal protection for LGBT people, especially for economic and healthcare matters, due to the high risk of violence and death for people who come out. Spanish tweets also talk about families with same-sex parents. Portuguese tweets show homotransphobic content and negative views on LGBT healthcare (61% negative sentiment).

4.4 Cultural representation

This category appears only in French, Italian, Portuguese, and Spanish.

Representation in media The tweets about LGBT representation in the media mainly feature in Italian and Spanish, and mostly focus on gay actors, characters and authors, often discussing their coming out. Although users are supportive of gay celebrities, they express negative sentiment (57% and 53% for Italian and Spanish respectively) due to the discrimination they faced.

Anti-LGBT language in sports The sentiment of discussions about sports is mostly negative (69% for French, 63% for Portuguese, and 48% for Italian). Homotransphobic slurs are frequently used to insult soccer and rugby players who perform poorly: this reflects the cultural association of masculinity with physical strength and athletic ability in these cultural contexts.

5 Discussion

Through our research, we have gained insight into the widespread use of homotransphobic language in all the languages we examined: despite hate speech detection systems are implemented, our findings suggest that there remains a significant amount of homotransphobic language. This highlights the pervasive nature of this issue and underscores the need for more targeted efforts to combat this phenomenon.

We found significant differences across languages. For instance, we found that in Norwegian, the derogatory term "f*ggot" ("bøg"), appeared in

only eight tweets across the entire dataset. This stands in stark contrast to the other languages we studied, where derogatory terms were more prevalent. It is clear that addressing this issue requires approaches that account for these cultural differences. Our findings have shed light on the higher incidence of homotransphobic language in religious and conservative cultural contexts, specifically in French and Italian tweets. We observed a link of this trend to misinformation, particularly to health issues such as monkeypox and vaccines. In addition, we observed the effects of politics on homotransphobic language: countries with less comprehensive LGBT-safety legislation had higher rates of such language use, underscoring the importance of effective frameworks to protect LGBT rights.

Interestingly, we found that derogatory language tends to be directed more frequently toward transgender rather than homosexual individuals in some of the languages, such as Spanish. This highlights the need for interventions that specifically address this issue, rather than using a broad approach.

6 Conclusion

We conducted a cross-lingual analysis of seven languages, examining how public discourse on Twitter frames LGBT individuals and issues. Our findings indicate that homotransphobic language continues to be prevalent despite the implementation of automatic hate speech detection models. Additionally we contributed a taxonomy for categorizing homotransphobic discourse, which can serve as a valuable tool to create datasets, as well as defining LGBT-related topics for analysis. By shedding light on the ways in which different cultures and languages frame LGBT issues, we hope that our study will contribute to ongoing efforts to promote acceptance and equality for all individuals.

Ethics statement

Similarly to Kennedy et al. (2022), we recognize that our analysis involved the examination of data containing a significant amount of hateful speech, which can be emotionally taxing and distressing for annotators. To address this concern, we provided our annotators with comprehensive information about the task's nature and the language and content they would encounter.

Furthermore, we took measures to ensure that the data we utilized for our analysis was gathered and utilized ethically and responsibly. We de-identified the data by eliminating tweet ids, user ids, and location data, utilizing only the raw text to guarantee that no personal data was accumulated or employed in any manner.

Limitations

We acknowledge that there exist numerous languages that may present distinctive challenges and characteristics regarding homotransphobia, beyond those examined in this paper. Our decision on which languages to include was based on various factors, including the accessibility of native speaker annotators, the global prevalence of each language, and the cultural and linguistic diversity they represent. Our dataset encompasses languages spoken worldwide, such as English, Spanish, Portuguese, and French, as well as more geographically specific languages, such as German, Italian, and Norwegian.

Our cross-linguistic comparison proved challenging due to the varying ratios of terms used in each language. For instance, we found that compared to other languages, Italian does not contain slurs directly targeting lesbian individuals.[5] Moreover, it presents more slurs with sexual connotation towards homosexual men. It is also important to note that personal experiences and exposure to certain types of language may influence the selection of keywords by native speakers, potentially skewing the distribution for some languages and introducing a strong sampling bias. To partially address this limitation we recruited, where possible, multiple native speakers per language, from diverse backgrounds.

Moreover, it should be noted that this study may not have fully captured the rich diversity of each language due to the possible exclusion of regional or dialectal differences that were not incorporated into the dataset. To partially address this limitation, we requested native speaker annotators to provide keywords that encompassed culturally-specific meanings that may not have direct translations in other languages. Nevertheless, obtaining a more comprehensive coverage of dialectal phrases for each language would have necessitated a larger number of annotators.

This is particularly apparent in the case of languages such as Spanish and Portuguese, which are official languages in both Southern Europe and

[5] https://www.gay.it/parole-insulto-lesbiche

Latin America. For instance, a word that is deemed to be homotransphobic in a Latin American country may not be considered offensive in Europe. To adequately address these variations in meaning and usage, a more nuanced approach would be necessary, which would entail dividing tweets by geographic location. While this avenue of research presents exciting possibilities for future studies, it would also entail additional challenges, such as the need for a larger and more diverse set of annotators to cover the different regions and dialects.

Acknowledgments

We thank the anonymous reviewers for their useful feedback, as well as Matyáš Boháček for insightful discussions on this topic. We would also like to thank the annotators who took part in this project: Benjamin Aston, Sergio Calo, Anaïs Giegerich, Costanza Moroni, Marie Pechenard, Ariadna Quattoni, Kilian Rothmund, and Samia Touileb. This project has partially received funding by Fondazione Cariplo (grant No. 2020-4288, MONICA). Greta Damo and Debora Nozza are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. Davide Locatelli is part of the INTERACT group of the Technical University of Catalonia, and is supported by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant No. 853459). We gratefully acknowledge the computer resources at Artemisa, funded by the European Union ERDF and Comunitat Valenciana, and the technical support provided by the Instituto de Fisica Corpuscular, IFIC (CSIC-UV).

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

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.

Gianmarco Biancalani, Lucia Ronconi, and Ines Testoni. 2022. Differences in social networking behaviors between italian gay and heterosexual men. *Sexuality Culture*, 27.

Claudia Bianchi. 2014. Slurs and appropriation: An echoic account. *Journal of Pragmatics*, 66:35–44.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

Paula Carvalho, Bernardo Cunha, Raquel Santos, Fernando Batista, and Ricardo Ribeiro. 2022. Hate speech dynamics against African descent, Roma and LGBTQI communities in Portugal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2362–2370, Marseille, France. European Language Resources Association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Biraj Dahal, Sathish A. P. Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(1):24.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2018)*, 12:59.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Ilse Hartmann-Tews, Tobias Menzel, and Birgit Braumüller. 2021. Experiences of lgbtq+ individuals in sports in germanyerfahrungen von lsbtq+-personen im sport in deutschland. *German Journal of Exercise and Sport Research*, 52.

Elise Margrethe Vike Johannessen. 2021. Blurred lines: The ambiguity of disparaging humour and slurs in norwegian high school boys' friendship groups. *YOUNG*, 29(5):475–489.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chaïm la Roi and Jornt J. Mandemakers. 2018. Acceptance of homosexuality through education? investigating the role of education, family background and individual characteristics in the united kingdom. *Social Science Research*, 71:109–128.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Ernesto Lee, Furqan Rustam, Patrick Bernard Washington, Fatima El Barakaz, Wajdi Aljedaani, and Imran Ashraf. 2022. Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcr-nn model. *IEEE Access*, 10:9717–9728.

Joanne Chen Lyu, Eileen Le Han, and Garving K Luli. 2021. Covid-19 vaccine–related discussion on twitter: Topic modeling and sentiment analysis. *J Med Internet Res*, 23(6):e24435.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258–264, Dublin, Ireland. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Debora Nozza and Dirk Hovy. 2022. The state of profanity obfuscation in natural language processing. *arXiv preprint arXiv:2210.07595*.

Jacob Pousher and Nicholas Kent. 2020. *The Global Divide on Homosexuality Persists: But Increasing Acceptance in Many Countries Over Past Two Decades*. Pew Research Center.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Hilde Slaatten, Norman Anderssen, and Jørn Hetland. 2015. Gay-related name-calling among norwegian adolescents – harmful and harmless. *Scandinavian Journal of Psychology*, 56(6):708–716.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Jia Xue, Junxiang Chen, Chen Chen, Chengda Zheng, Sijia Li, and Tingshao Zhu. 2020. Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PLOS ONE*, 15(9):1–12.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

A Data

A.1 Keywords

Warning: Because obfuscated words are often not discernible, especially for non-native speakers (Nozza and Hovy, 2022), here we reported the keywords in their unobfuscated form. This section contains keywords readers may find upsetting and offensive.

German schwul (*queer*), schwuler (*queer*), lesbe / lesbo (*lesbian*), schwuchtel (*faggot*), schwanzlutscher (*cocksucker*), transe (*tranny*), tunte (*faggot*), schwuppe (*queer*), homo (*homosexual*), standgebläse (*short cocksucker*), tucke (*faggot*), schwulette (*faggot*), butch (*butch lesbian*), lesbich (*lesbian*), gay (*gay*), schranklesbe (*closeted lesbian*).

English sissy, fag, faggot, poof, cocksucker, homosexual, lesbo / lesbian, genderbender, dyke, transvestite, sodomite, gay, cuntboy, ladyboy, tranny / trannie, genderfuck, fudgepacker.

Spanish homosexual (*homosexual*), maricón / marica (*fag*), amanerado / a (effeminate), lesbiana (*lesbian*), trolo (*fag*), guey / guei / gay (*gay*), desviado (*deviate*), sodomita (*sodomite*), marimacho / marimacha (*butch lesbian*), sarasa (*fag*), travelo (*tranny*), joto (*faggot*), travestido (*transvestite*), soplanucas (*fudgepacker*), muerdealmohadas (*assfucked*), safista (*lesbian*).

French enculé (*assfucked*), homosexuel(le) (*homosexual*), transgenres (*transgender*), fiotte (*faggot*), tapette (*fag / fly swatter*), lopette (*sissy*), folle (*crazy woman, or gay queen, in slang*), pédale (*faggot*), balasko (*butch lesbian*), tarlouze (*poof*), tafiole (*faggot*), pédé(raste) / PD (*homosexual male*), fif (*effeminate gay*), gouine (*dyke*), tantouse (*faggot*), lesbienne (*lesbian*).

Italian gay (*gay*), pride (*pride*), lesbica (*lesbian*), frocio (*queer*), finocchio (*faggot*), ricchione (*faggot*), checca (*effeminate gay*), succhiacazzi (*cocksucker*), culattone (*fudgepacker*), rottinculo (*assfucked*), piglianculo (*assfucked*), effeminato (*effeminate*), bocchinaro (*cocksucker*), pompinaro (*cocksucker*), travione (*tranny*).

Portuguese homossexual (*homosexual*), viado (*faggot*), bicha (*faggot*), maricas (*faggot*), transexual (*transexual*), fufa (*dyke*), panasca (*faggot*), larilas (*faggot*), panilas (*faggot*), panaleiro (*faggot*).

Norwegian skeiv (*queer*), transkvinne (*trans woman*), transperson (*trans person*), homse (*homo*), transkjønnet (*transgender*), bifil (*bisexual*), transmann (*trans man*), soper (*faggot*), dyke (*dyke*), transe (*tranny*), lesbe (*lesbian*), bøg (*faggot*), homo (*homo*), kuksuger (*cocksucker*), rompis (*fudgepacker*), skinkerytter (*fudgepacker*), gay (*gay*).

L	COLLECTED	ESTIMATED
DE	44,889	314,082
EN	1,070,280	31,886,162
ES	164,451	2,003,997
FR	93,395	1,103,618
IT	59,830	1,021,508
NO	5,036	14,777
PT	38,070	2,343,635

Table 4: Estimate of number of tweets posted in the week 06/01-07/2022 by language (**L**), along with the number of tweets we collected containing the LGBT keywords.

A.2 Collection and processing

We cleaned our data by removing stopwords. We used the stopword lists available at https://github.com/stopwords-iso/stopwords-iso. Additionally we removed duplicates, mentions, hashtags, and URLs. To speed up the analysis, we randomly sampled 25,000 tweets from each language, except for Norwegian, which had fewer tweets. We checked that our samples were similar to the original data by comparing the frequency of each keyword in both datasets.

To investigate why there were fewer Norwegian tweets, we sought to determine whether this was due to a lower overall volume of tweets from Norwegian users. To do this, we selected commonly used words in each language (specifically, "I", "you", "say", and "think") and we tallied the number of tweets containing these words in the week of 06/01-07/2022 using the Postman API Network[6], as a proxy for each language's tweet volume. Our analysis revealed that the average number of weekly tweets in Norwegian was considerably lower than that of the other languages. Therefore, the lower number of gathered Norwegian tweets was not due to a lack of Norwegian individuals tweeting about LGBT issues, but rather a general trend of lower tweet volume in the language. We present our language-specific tweet counts in Table 4.

A.3 Data Statement

We follow Bender and Friedman (2018) and provide a Data Statement for the collection of tweets we used in our study.

[6] https://www.postman.com/

Curation rationale The goal of our project was to collect a large and multilingual collection of tweets relevant to LGBT issues, and characterize the differences in public discourse around these topics in the different linguistic contexts. For this purpose, we employed a team of native-speakers to devise a list of keywords that could be used to search posts with Twitter's historical API. Our data points consist of tweet IDs and the raw text of the tweet. We do not provide labels that accompany the text. Due to the nature of the research, a large proportion of the data we collected contains hurtful and/or explicit messages.

Language variety Our data covers seven languages: German, English, French, Italian, Norwegian, Spanish, and Portuguese.

Annotator demographics The keyword selection has been done by a group of ten native speakers belonging to the 25-35 age group, all with experience in computational linguistics and familiar with LGBT issues. The taxonomy has been developed by two annotators in the 25-35 age group, in a multi-round process involving also the labeling of topics. Both annotators are experienced in computational linguistics and LGBT issues. Because the two annotators are not native speakers of all the languages involved in the project, their annotation has been aided with an automatic translation software.

Speech situation All data was obtained using the Twitter's historical API and consists of tweets that appeared on the platform between 05/01/2022 and 09/01/2022.

B Experimental setup

B.1 Methodology

Topic Modeling Within CTM, we used a distilled multilingual Universal Sentence Encoder (Yang et al., 2020) from the sentence-transformers library (Reimers and Gurevych, 2019) to encode sentences into vectors. We trained the model for 10 epochs and tested it with 5, 10, 15, and 20 topics. We used the NPMI score (Lau et al., 2014) to assess the coherence of the topics. We found that 10 topics were optimal for most languages (see Figure 1).

Sentiment Analysis We classified each tweet as negative, neutral, or positive using a pretrained sentiment analysis model (Barbieri et al., 2022). The model is fine-tuned on tweets and can interpret emotions across different languages. While it is not fine-tuned on every languages, the authors demonstrate that the model has good generalization capabilities to unseen languages.

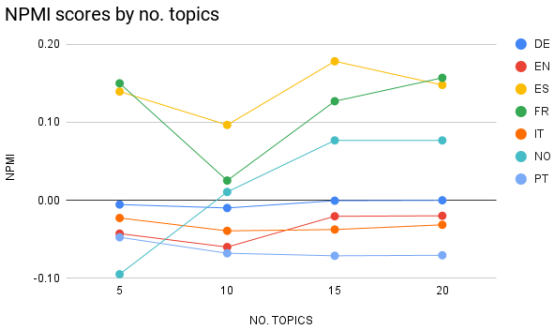

Figure 1: NPMI scores by number of topics for each language (lower is better). We can observe that the score is lowest for 10 topics for all languages, with the exception of Norwegian.

Because Norwegian is not among the training languages, we further investigate to convalidate the results of XLM-T (Barbieri et al., 2022) for sentiment analysis in Norwegian. We compared the sentiment scores on automatic English translations of Norwegian tweets to the scores on the original text. The results were similar, indicating reliable results for all languages. We illustrate them in Figure 2.

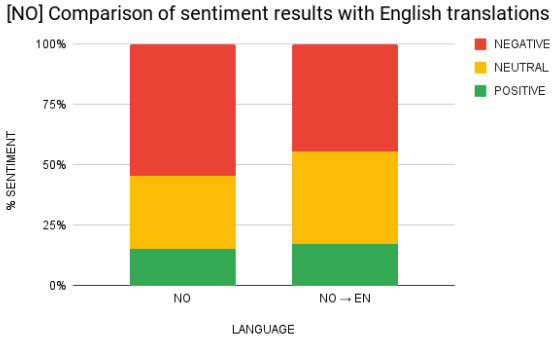

Figure 2: Comparison of sentiment analysis results on original Norwegian tweets (NO) versus automatic English translations of the tweets (NO → EN).