ISSN 1870-4069

Dimensionality reduction based on Chi-Square Statistic and Testors for LGBT+phobia Detection

Metztli Ramírez-González, Jesús Ariel Carrasco-Ochoa,
José Francisco Martínez-Trinidad

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Mexico

{metztli.ramirez, ariel, fmartine}@inaoep.mx

Abstract. LGBT+phobia detection is a text classification task allowing identifying this kind of discrimination. However, text classification is a challenge due to the high dimensionality of word representations. In this work, we propose combining Chi-Square (χ^2) Statistic and irreducible testors to reduce dimensionality in LGBT+phobia Detection. The results in a public database (specially built for hate speech detection tasks) indicate that our proposal allows obtaining a small representation space to reach as good classification results as using the whole representation space's dimensionality for classifying and detecting LGBT+phobia in Mexican Spanish.

Keywords: LGBT+phobia detection, feature selection, testors.

1 Introduction

LGBT+phobia refers to any discrimination based on sexual preferences and/or gender identity [5], previously generalized as Homophobia [6]. According to data reported by the National Discrimination Survey (Enadis) 2022, more than 3.3 million people in Mexico report or have reported a non-normative sexual orientation or gender identity, representing approximately 3.6% of the national population. However, discrimination and violence towards the LGBT+ community do not occur in isolation; Enadis itself reports discrimination regarding access to housing, public office, health services, and even family rejection [7]. The CNDH [4] denounces that the LGBT+ community is often the victim of harassment, torture, arbitrary detention, and even murder, many times with total impunity.

According to the Mexican government [10], in the period from 2012 to 2022, Conapred has registered 1,175 complaints related to sexual and gender diversity. Moreover, according to Forbes [18], Mexico is the second country in Latam with the most hate crimes against the LGBT+ community. In Mexico, from 2019 to 2022, the LGBT+ community has been the victim of at least 305 violent acts motivated by hate, including murders, disappearances, and attempted murders, among others. Even so far, in 2024, at least 25 cases of murder against LGBT+ people have been reported [8].

These figures make this problem not only essential but also critical. These acts of discrimination find a space on social media, promoting and normalizing violence through hateful comments. Therefore, promptly detecting LGBTQ+phobic messages can improve the content moderation and create safer online environments [13].

LGBT+phobia can be included within the so-called hate speech. These messages are difficult to identify as they are influenced by various aspects such as the domain of a statement, its discursive context, concurrent media objects (images, videos, and audio), the historical and world context, and the identity of the author and recipient [21]. In recent years, research into hate speech detection has gained momentum, leading to exploring various techniques to address this problem.

Traditional methods typically leverage term frequency representations, which have shown promising results when combined with conventional classification approaches [23]. However, these representations produce high-dimensional spaces, which are difficult for classifiers to handle. Thus, the most recent solution approaches focus on the use of Transformers [9, 15–17, 20, 22, 26]; the problem with these approaches is that they often lack explainability.

To face the problems of explainability and high-dimensional representation spaces, in this work, this work proposes combining Chi-Square (χ^2) Statistic and irreducible testors [14] to reduce dimensionality in LGBT+phobia Detection. Instead of a transformer approach, we use the vocabulary of the problem domain and the frequency of terms in it. Our results in a public database show that our proposal can reduce the representation space while getting as good results as those by using the whole representation space's dimensionality for the classification and detection of LGBT+phobia in Mexican Spanish.

This work is based on the work developed in [19], where an alternative idea for addressing the identification of LGBT+phobia for the shared task HOMO-MEX 2024 for IberLEF [3, 12, 13] is proposed.

This paper is organized as follows: Section 2 summarizes the solutions given for the HOMO-MEX 2023 edition. Section 3 describes the proposed solution. Section 4 describes the obtained results. Finally, in section 5, we provide our conclusions and some directions for future work.

2 Related Work

HOMO-MEX is the first shared task focused on detecting LGBT+phobia in Mexican Spanish, organized for the first time in 2023 [1]. Mexican Spanish is characterized by its richness in language, such as the use of metaphors, allegories, figures of speech, insults, and nicknames that require a lot of context to be understood.

The proposals seen at Homo-Mex 2023 involved using Transformer-based models and data augmentation techniques [19]. Shahiki-Tash et al. [22] point out the importance of performing text preprocessing before using classification models so that they have better performance. Moriña et al. [17] used different

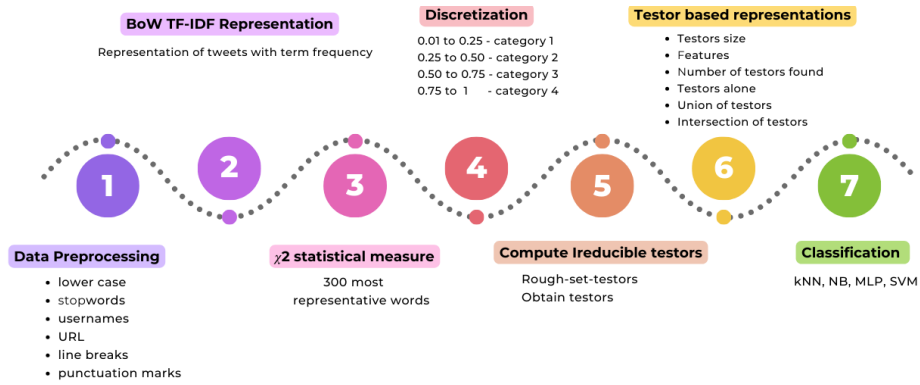

Fig. 1. Proposed solution diagram.

Transformer models and compared their performance. Marrugo-Tobón et al. [16] and Yigezu et al. [26] used data augmentation techniques and different variants of BERT for classification. García-Díaz et al. [9] experimented with combinations of Spanish and multilingual large language models (LLMs). Rosauro and Cuadros [20] compared traditional Classification Models and Transformers. Finally, Macias et al. [15] performed LGBT+phobia classification with different text representations based on frequency, without applying feature selection techniques and using classic models such as SVM and Bagging.

Most of these proposals focus on the use of Transformers and do not address feature selection, which is essential to characterize the problem domain. In this proposal, we focus on the frequency of the terms present in the problem vocabulary, applying dimensionality reduction based on χ^2 and irreducible testors to better understand the problem space.

3 Proposed Solution

Our approach is based on first applying a feature selection step over the BoW with TF-IDF representation using χ^2, and then a second feature selection step using irreducible testors. Figure 1 shows a schematic representation of the proposed approach, which consists of seven stages, we describe each of these stages in the following sections.

Data Preprocessing: All tweets are converted to lowercase and preprocessed by removing usernames, punctuation marks, URLs, line breaks, and stopwords.
BoW TF-IDF Representation: To represent the tweets, we used the classic Bag of Words (BoW) TF-IDF representation, which accounts for the term frequency in the document collection. Only those terms appearing in at least ten tweets were considered.
Chi-Square (χ^2) Statistic: From the BoW TF-IDF representation, the most relevant terms for class identification are calculated using the chi-square χ^2

statistic. The 300 terms with the highest statistical values are selected to represent the data.

Discretization: With the 300 selected terms, for allowing computing irreducible testors, a new BoW TF-IDF representation is built. Based on this representation, values are discretized into the following ranges: 0.01 to 0.25 - category 1, 0.25 to 0.5 - category 2, 0.5 to 0.75 - category 3, 0.75 and above - category 4.

Compute Ireducible testors: The RCC-MAS algorithm is applied on the discretized representation to find some irreducible testors. A testor is a subset of features that can distinguish objects from different classes in a data set, i.e., no object of one class is confused with any object of another class [14]. In our proposed approach, a testor becomes a set of words selected based on their frequency in the data. These words constitute a sufficient vocabulary to represent the data and allow us to differentiate between the classes. RCC-MAS is an algorithm designed to calculate all constructs; however, we employed it according to [11] to compute irreducible testors, but given that the search space for searching irreducible testors is very large, the algorithm is given a fixed execution time to just finding some testors.

Testor-based representations: The obtained irreducible testors are analyzed to determine their characteristics, such as the number of irreducible testors found, the words they contain, and their size. To evaluate the irreducible testors, the data representation is modified so that the BoW TF-IDF, previously reduced to 300 words, is characterized only by the words included in a testor, and then classification is performed. Additionally, representations and classifications are evaluated with the intersection and union of the identified irreducible testors. Finally, the results obtained using the 300 words are compared with those obtained using the words identified by the irreducible testors.

Classification: From each testor-based representation, the classification is performed employing k-Nearest Neighbors with k=10, Naive Bayes, Multilayer Perceptron, and SVM, using 5-fold cross-validation. The classification quality was measured using F1-score.

4 Experiments

4.1 Dataset

To evaluate the methodology we use the dataset provided by HOMO-MEX. HOMO-MEX is the first corpus for the detection of LGBT+ phobia in Mexican Spanish. It comprises public tweets extracted using the Twitter API, including keywords used in LGBT+phobic contexts. First, a list of nouns used to refer to the LGBT+ community was compiled (see Figure 1). Then, more than ten thousand tweets containing any of these nouns from the last ten years were selected. Four annotators subsequently labeled each tweet as LGBT+phobic, non-LGBT+phobic, or not related to the LGBT+ community [25]. HOMO-MEX hate speech detection task [3, 5, 12]: has a total of 8800 training data, divided into 5482 instances for the Non-LGBT+phobic class, 1072

Table 1. LGBT+phobia detection results on the different representations used in our experiments (F1-score).

		Classifiers			
Representation	**Size in words**	**SVM**	**kNN**	**NB**	**MLP**
BoW- All words	1191 words	**0.833332**	0.803367	0.775447	0.823568
300 words	300 words	0.819405	0.801503	0.681794	**0.823116**
Testor 1	133 words	0.803647	0.790423	0.458807	0.801504
Testor 2	133 words	0.803909	0.792035	0.460065	0.803985
Testor 3	134 words	0.805923	0.797587	0.460802	**0.811365**
Testor 4	134 words	0.805466	0.793303	0.450702	0.810714
Testor 5	134 words	0.802722	0.789211	0.455831	0.804951
Testor 6	134 words	0.802229	0.79703	0.452902	0.804168
Union	137 words	0.806558	0.798218	0.461325	0.803276
Intersection	131 words	0.802903	0.799502	0.459348	0.804686

instances for the LGBT+phobic class, and 2246 instances for the irrelevant class.

HOMO-MEX has several subtasks, but this work focuses on LGBT+phobic speech detection, which aims to predict the label of each tweet. It is a multiclass task in which a tweet can belong to one of the next three classes:

- LGBT+phobic (P), which includes tweets that contain hate speech directed against persons whose sexual orientation and/or gender identity differs from heterosexuality.
- Non-LGBT+phobic (NP), which includes tweets that mention concepts related to the LGBT+ population but without any intention of hate speech.
- Non-LGBT+ related tweets (NR) those that have no relationship with the LGBT+ community.

4.2 Application of the Proposed Solution

Six different testors were obtained by applying the proposed methodology and the RCC-MAS algorithm. Testors 1 and 2 contain 133 words each, while Testors 3, 4, 5, and 6 contain 134 words each. Combining these testors (union) results in 137 words, and the intersection results in 131 words. Table 1 shows the results of the experiments on LGBT+phobia detection (in terms of F1-score) using the entire vocabulary, the 300 most representative words, and the testor-based representations, including their union and intersection.

4.3 Analysis of Results

Some highlights found after analyzing the results of our experiments are:

Best Results: The best results were obtained using the entire vocabulary without attribute selection (1191 words). However, the results obtained with the initial choice of 300 words (F1-score of 0.823116) and the 134-word testors (F1-score of 0.811365) are not very different from the best result (F1-score of 0.8333). Applying the Kruskal-Wallis test, a p-value of 0.437 is obtained for all classifiers, indicating that there are no statistically significant differences between the different word representations. Furthermore, according to the t-test, there is no statistical difference when comparing the use of all the terms against the 300-word selection by chi-square statistic(χ^2) or against the testor-based representations. Additionally, the results of the Nemenyi test do not show any significant variation when comparing pairs of representations. Therefore, we can conclude that there are no significant differences between all the representations tested. This suggests that the 134 words selected by the irreducible testor three are sufficient to adequately represent the data and provide acceptable classification performance. It is also important to highlight that words contained in the six testors are almost the same, with a variation of only 6 words. Thus, we can conclude that the vocabulary sufficient to characterize the entire data set is about 131 words.

Selected Words: The 137 words chosen by at least one irreducible testor are listed below:
'persona', 'putas', 'nomás', 'transexual', 'si', 'mariconcito', 'acuerdo', 'maría', 'borracho', 'loquita', 'personas', 'estupido', 'bis', 'facewithrollingeyes', 'gays', 'comunidad', 'mujeres', 'jajajajaja', 'mariquita', 'acá', 'cis', 'hoy', 'feministas', 'maricas', 'loveislove', 'discriminación', 'paisanos', 'show', 'hombres', 'editorial', 'clóset', 'llorar', 'parejas', 'at', 'puto', 'lucha', 'gay', 'feminista', 'maricones', 'identidad', 'gente', 'lea', 'sinónimo', 'liosas', 'nombre', 'problema', 'sánchez', 'mismo', 'sexo', 'vida', 'maricon', 'género', 'pareja', 'súper', 'volviendo', 'sorprende', 'ser', 'vato', 'mariquitas', 'historia', 'queer', 'bi', 'in', 'veces', 'pues', 'chiva', 'campeonato', 'puro', 'lesbiana', 'puros', 'primera', 'aclaro', 'diciendo', 'putito', 'musicalnotes', 'hablar', 'negras', 'solo', 'pinche', 'santa', 'transformer', 'ke', 'hetero', 'homosexuales', 'vergas', 'respeto', 'tragedias', 'bisexual', 'bisexuales', 'loca', 'dragas', 'pinches', 'jotos', 'vestidas', 'heterosexuales', 'putos', 'madre', 'bandera', 'enseñando', 'joto', 'maricón', 'paso', 'méxico', 'hombre', 'chicos', 'lesbianas', 'travesti', 'tema', 'the', 'gayboy', 'transexuales', 'gol', 'aguantas', 'mujer', 'locas', 'drag', 'marica', 'odiantes', 'puta', 'raritos', 'vuelta', 'sexual', 'jotito', 'vestida', 'homosexual', 'transgénero', 'transformers', 'rarita', 'trans', 'derechos', 'puñal', 'jajaja', 'papel', 'draga', 'calor', 'puñetas'.

4.4 Significance of Words by Class

The most significant words according to the Chi-Square (χ^2) Statistic for each class are shown below:

– **Non-LGBT+phobic class:** trans, gay, gente, mujeres, homosexual, homosexuales, lesbiana, mujer, drag.

- **LGBT+phobic class:** joto, marica, jotos, maricon, puto, mariquita, maricas, gay, puñal, pinche.
- **Non-LGBT+ relate Class:** loca, vestida, puta, puto, bi, locas, transformers, hoy, bis, madre.

This analysis of the vocabulary used in each class shows marked differences, especially in the presence of insults and derogatory words in the LGBT+phobic class, in contrast to the other two classes that have fewer terms of this type. This language analysis can help identify and build a lexicon-based solution.

5 Conclusions

In this paper, we present a different approach to detecting LGBT+phobia in Mexican Spanish, by applying dimensionality reduction techniques, first a reduction to 300 vocabulary words selected using chi-square χ^2, and a further reduction using irreducible testors, obtaining a space of 134 words that seems sufficient to represent the whole dataset. LGBT+phobia detection results are reasonable compared to using other approaches, and this approach based on dimensionality reduction techniques has the property of knowing which specific words or terms we are representing the LGBT+ phobic tweets, which is lost if we apply the approach based on transformers, therefore, with the proposed approach in this paper, we contribute explainability to the problem.

We can conclude that linguistic analysis is a great tool for understanding the problems related to hate speech, and this approach can be beneficial and even improve some of the current techniques. In future work, we will evaluate the usefulness of constructs [24] and Goldman fuzzy reducts [2] from rough set theory, which can work directly on the BoW TF-IDF representation without discretizing it.

Acknowledgments. This research was partially supported by the National Council of Humanities, Sciences, Technologies, and Innovation of Mexico (CONAHCyT) through its graduate study scholarship program.

References

1. Bel-Enguix, G., Gómez-Adorno, H., Sierra, G., Vásquez, J., Andersen, S.T., Ojeda-Trueba, S.: Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish-speaking lgbtq+ population. Natural Language Processing 71, 361–370 (2023)
2. Carrasco-Ochoa, J., Lazo-Cortés, M., Martínez-Trinidad, J.: An algorithm for computing goldman fuzzy reducts. In: Pattern Recognition: 9th Mexican Conference, MCPR 2017, Huatulco, Mexico, June 21-24, 2017, Proceedings. Lecture Notes in Computer Science, vol. 9, pp. 3–12. Springer, Cham (2017)
3. Chiruzzo, L., Jiménez-Zafra, S.M., Rangel, F.: Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages.

In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org (2024)
4. CNDH: Día internacional contra la homofobia, la transfobia y la bifobia. https://www.cndh.org.mx/noticia/dia-internacional-contra-la-homofobia-la-transfobia-y-la-bifobia-0, (accessed 01 July 2024)
5. CodaLab: HOMO-MEX: Hate speech detection towards the Mexican Spanish speaking LGBT+ population. https://codalab.lisn.upsaclay.fr/competitions/10019, (accessed 01 July 2024)
6. CONAPRED: Guía para la acción pública contra la homofobia. México, D. F. (2012)
7. CONAPRED: Discriminación en contra de las personas por su orientación sexual, características sexuales e identidad y expresión de género. México, D. F. (2024)
8. FundaciónArcoiris: Observatorio nacional de crímenes de odio contra personas lgbt. http://www.fundacionarcoiris.org.mx/agresiones/panel, (accessed 02 July 2024)
9. García-Díaz, J.A., Jiménez-Zafra, S.M., Valencia-García, R.: Umuteam at homo-mex 2023: Finetuning large language models integration for solving hate-speech detection in mexican spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
10. GobiernoMéxico: Registra conapred mil 175 quejas relacionadas con personas de la diversidad sexual y de género. https://www.gob.mx/segob/prensa/registra-conapred-mil-175-quejas-relacionadas-con-personas-de-la-diversidad-sexual-y-de-genero, (accessed 01 July 2024)
11. González-Díaz, Y.: Rcc-mas. GitHub. https://github.com/ygdiaz1202/RCC-MAS/blob/master/out/artifacts/RCC_MAS_jar/RCC-MAS.jar, (accessed 05 June 2024)
12. Gómez-Adorno, H., Bel-Enguix, G., Calvo, H., Vásquez, J., Andersen, S., Ojeda-Trueba, S., Alcántara, T., Soto, M., Macias, C.: Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population. Natural Language Processing 73 (2024)
13. HOMO-MEX: Homo-mex 24: Hate speech detection towards the mexican spanish speaking lgbt+ population. https://sites.google.com/view/homomex/home, (accessed 01 July 2024)
14. Lazo-Cortes, M., Ruiz-Shulcloper, J., Alba-Cabrera, E.: An overview of the evolution of the concept of testor. Pattern Recognition 34(4), 753–762 (2001)
15. Macias, C., Soto, M., Alcántara, T., Calvo, H.: Impact of text preprocessing and feature selection on hate speech detection in online messages towards the lgbtq+ community in mexico. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
16. Marrugo-Tobón, D.A., Martinez-Santos, J.C., Puertas, E.: Natural language content evaluation system for multiclass detection of hate speech in tweets using transformers. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
17. Moriña, A.J.M., Pásaro, J.R., Vázquez, J.M., Álvarez, V.P.: I2c-uhu at iberlef-2023 homomex task: Ensembling transformers models to identify and classify hate messages towards the community lgbtq. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)

18. México, F.: Pese a los avances legales, méxico lidera en crímenes de odio contra personas lgbt. https://www.forbes.com.mx/mexico-lidera-crimenes-odio-personas-lgbt-avances-legales/, (accessed 02 July 2024)
19. Ramírez-González, M., Hernández-Farías, D., Montes-y Gómez, M.: Labtl-inaoe at homo-mex 2024: Distance-based representations for lgbt+ phobia detection. In: XL Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2024) (2024), (In press)
20. Rosauro, C.F., Cuadros, M.: Hate speech detection against the mexican spanish lgbtq+ community using bert-based transformers. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
21. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media. pp. 1–10 (2017)
22. Shahiki-Tash, M., Armenta-Segura, J., Ahani, Z., Kolesnikova, O., Sidorov, G., Gelbukh, A.: Lidoma at homomex2023@iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
23. Shridhara, M., Pristaš, V., Kotvytskiy, A., Antoni, L., Semanišin, G.: A short review on hate speech detection: challenges towards datasets and techniques. In: Proceedings of the 2023 World Symposium on Digital Intelligence for Systems and Machines (DISA). pp. 204–209. IEEE (2023)
24. Susmaga, R.: Reducts versus constructs: an experimental evaluation. Electronic Notes in Theoretical Computer Science 82(4), 239–250 (2003)
25. Vásquez, J., Andersen, S., Bel-Enguix, G., Gómez-Adorno, H., Ojeda-Trueba, S.L.: Homo-mex: A mexican spanish annotated corpus for lgbt+ phobia detection on twitter. In: Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH). pp. 202–214 (2023)
26. Yigezu, M.G., Kolesnikova, O., Sidorov, G., Gelbukh, A.: Transformer-based hate speech detection for multi-class and multi-label classification. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)