# Knowledge Expansion Guided by Justification for Improved Sexism Categorization

Notebook for the EXIST Lab at CLEF 2025

Kapioma Villarreal-Haro[1,*,†], Fernando Sánchez-Vega[1,2,†] and Adrián Pastor López-Monroy[1,†]

[1] Computer Science Department, Mathematics Research Center (CIMAT), Jalisco S/N Valenciana, 36023, Guanajuato, Guanajuato, México

[2] Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Av. Insurgentes Sur 1582, Col. Crédito Constructor, 03940, CDMX, México.

**Abstract**

We describe in this paper the participation of team *CIMAT-GTO* in Task 1 (hard label setting) of EXIST 2025, which focuses on identifying sexism in tweets, determining the source's intention, and categorizing the types of sexism expressed. We propose a hybrid methodology that combines generative large language models with fine-tuned transformer-based classifiers through knowledge expansion. Our approach utilizes a generative model to highlight contextually relevant elements for the task as well as provide classification answers, and subsequently extracts justification texts that support the given predictions. We then conduct a justification-guided knowledge expansion when fine-tuning a smaller transformer-based model for classification, aiming for the model to learn from the reasoning encoded in the generated texts. We evaluate both monotask and multitask fine-tuning strategies and implement ensemble methods to improve robustness. Our results demonstrate that knowledge expansion using justifications obtained from generative models enhances performance over baseline few-shot classification and fine-tuned models. The proposed systems prove to be competitive and achieve second place in all three textual tasks.

**Keywords**

Generative Large Language Models, LLM Reasoning, Knowledge Expansion, Sexism Detection, Sexism Categorization, Social Media

## 1. Introduction

Social media platforms have become a widely used medium for communication and information consumption in recent years. These platforms, despite allowing for quick and easy information exchange among users, lead to problems such as exposure to misinformation and biased content [1]. Among the different social health problems that emerge in these platforms as a reflection of real-life problems, sexism is one of the most concerning ones as it is deeply embedded in societal norms and cultural attitudes [2]. Both overt hostile and subtle forms of sexism have been studied to be recurrent and negatively affect the psychological well-being of people in everyday interactions [3]. This phenomenon translates from the physical world to digital spaces, where it follows its own dynamics that frequently amplify extreme viewpoints due to anonymity and online disinhibition effects [4]. To better shape and understand the problem, we need to study how technology transforms social interactions and the impact that it has from different perspectives: technology can be viewed as a facilitator for gender-based violence [5], but also as a tool to challenge it and raise awareness [6].

However, current detection systems struggle with contextual understanding and nuanced categorization. Understanding the presence and prevalence of this phenomenon, and its many-sided manifestations, sheds light on the necessity of developing systems that are capable of identifying and characterizing this type of content while also managing large volumes of information.

*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*

[*] Corresponding author.

[†] These authors contributed equally.

kapioma.villarreal@cimat.mx (K. Villarreal-Haro); fernando.sanchez@cimat.mx (F. Sánchez-Vega); pastor.lopez@cimat.mx (A. P. López-Monroy)

This paper describes CIMAT-GTO's participation in Task 1 of EXIST (sEXism Identification in Social neTworks) 2025 [7, 8]: identifying sexism in tweets, determining the source's intention, and categorizing the types of sexism expressed. We present a hybrid approach that utilizes fine-tuning of transformer-based models with knowledge expansion using contextualized reasoning produced with generative LLMs. We use *Gemini-1.5-Flash* due to its general-purpose capabilities, and relatively moderate parameter size. This choice leaves room for further experimentation with larger or specialized reasoner models that may provide higher-quality responses and justifications, potentially enhancing performance following our proposed framework.

## 2. Related Work

In previous editions of EXIST, transformer-based architectures, such as BERT or RoBERTa, were the most widely used and effective models in textual tasks. These models were typically fine-tuned with or without prior pre-training, combined with techniques such as data augmentation, hierarchical classification, annotator information injection, model cascades, and ensembles [9, 10, 11, 12]. Meanwhile, during the 2024 edition of EXIST, generative models were not only being explored as zero or few-shot classifiers, but also in hybrid approaches where their answers or knowledge were combined or leveraged by other LLMs. Still, they were unable to surpass other types of models proposed, despite their impressive performance [12]. While fine-tuning is a classical framework for classification, where the input data relies only on text, there is work where performance is improved using injection of external knowledge like handcrafted textual features, graphs of knowledge, or retrieval from outside sources to provide with an expanded range of information [13, 14]. Among several applications, utilizing knowledge extraction of rationales to be further processed by a transformer alongside the tweet and use the expanded input for classification has been shown helpful for hate speech detection [15] and sexism detection [16], demonstrating how outputs from generative models can be leveraged by smaller classifiers in this type of schemes.

## 3. The data and the tasks addressed

sEXism Identification in Social neTworks (EXIST) is a shared task that aims to foster sexism identification and characterization in social media [8, 7]. The most recent edition includes textual and multimodal challenges. In this work, we focus only on text data from tweets.

### 3.1. Data

The tweet dataset comprises 10,034 tweets in both Spanish and English. The dataset is mostly balanced across languages and is roughly split into a (70:10:20) proportion for training, development, and testing, maintaining proportions for each class.

Six annotators labeled each tweet, and hard labels were computed by organizers following probabilistic thresholds. As we focus on predicting hard labels, we only considered tweets where such categorization was retained. In the case of task 1.1, binary identification of sexism-related content, this reduced the set to approximately 87% of the original dataset.

### 3.2. The tasks

The three textual tasks addressed, in the hard label setting, in this work are

- Task 1.1: Sexism Identification: Binary classification to detect sexism-related content.
- Task 1.2: Source intention. Classification of intention if sexism-related content was identified. Three possible categories: direct, reported, and judgemental.

- Task 1.3: Sexism Categorization. Multi-label classification of sexism-related content was identified. Possible categories include ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence.

Tasks 1.2 and 1.3 derive from the classification of Task 1.1 and provide more insights into the characterization of sexism-related content.

## 4. Methodology

We propose hybrid systems that utilize generative LLMs to obtain answers to classification tasks *(FS-Ctx)*, and then extract justification texts that explain the given answers. We then use these texts as knowledge expansion to fine-tune transformer-based classifiers (*FT-KE-Mono* and *FT-KE-Multi*). For each of these fine-tuned variants, we train three models as an ensemble to improve robustness and performance (*Ens-FT-KE-Mono* and *Ens-FT-KE-Multi*). This two-stage process is detailed in the following subsections. An overall description of the individual systems is provided in Figure 1, while Table 1 summarizes the models studied and their corresponding run numbers.

### 4.1. Generation Stage

The first step in our methodology involves prompting an LLM to provide contextual cues about the tweet and to answer the three tasks within a single query. Items in the set of contextual elements are obtained by identifying recurrent patterns in the automatically discovered relevant information for the task obtained in the generation stage of the proposal of other studies [17]. This contextualization allows the model to retrieve a structured list of relevant elements aligned with the task and improves the accuracy of classification using generation. The prompt is also enriched with a few-shot examples and definitions for possible answer categories, as these are popular strategies to improve performance [18]. At this stage, the model chosen is *Gemini-1.5-Flash*, as it is a general-purpose model and offers a good trade-off between performance and efficiency.

***FS-Ctx*** **(Few-Shot-Context)** This is the generation-based method following the previously described setting. Although the generation following this methodology outperforms other prompt-based methods and some simple classifiers, it falls short compared to the best-performing classifier models.

We explore in the next stage how to achieve stronger models leveraging knowledge encoded by generative models that is indirectly present in the answers. In an intermediate step, we further prompt our generative LLM to justify the answers produced in the previous step based on the classification. We expect these texts to be informative and to enhance model performance, as they are used to enrich the input for classifier models.

### 4.2. Fine-Tuning Stage with Knowledge Expansion

We fine-tune two different types of models, using our proposed knowledge expansion strategy that concatenates justifications and the original tweet, and feeds these as input to the model, leveraging not only the classification preferences of the particular generative LLM from the previous step, but also reasoning in justifications. We evaluate two different types of fine-tuning techniques. We choose an XLM-RoBERTa trained on multilingual tweets [19] as a base model, to leverage information from both languages in the corresponding domain.

***FT-KE-Mono*** **(Fine-Tuned with Knowledge Expansion, Monotask)** Individual fine-tuning of the RoBERTa-based model for each subtask of task 1. Hence, we rely on three different models to provide a complete response to all text tasks. A post-processing correction process is applied to labels predicted in tasks 1.2 and 1.3 to align with the negative output of task 1.1 and avoid contradiction.

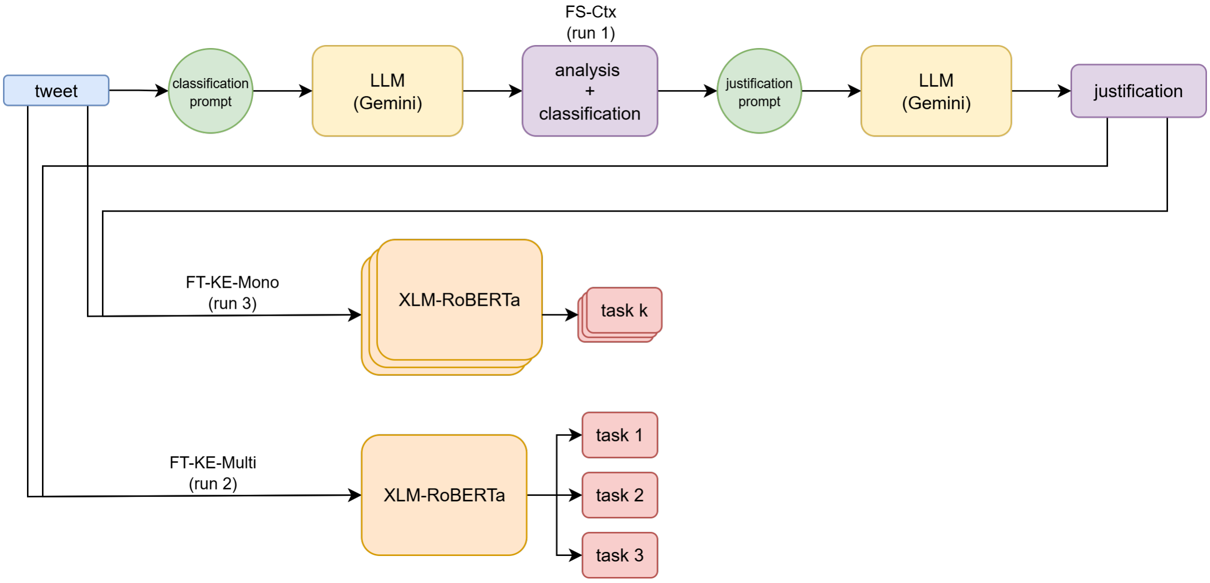

**Figure 1:** Overall description of system components and flow. *(FS-Ctx)* consists of prompt-based classification, and (FT-KE-Mono and FT-KE-Mono correspond to individual fine-tuned systems)

***FT-KE-Multi* (Fine-Tuned with Knowledge-Expansion, Multitask)** Multitask learning to predict labels for the three tasks at the same time, using a single multilingual RoBERTa-based model. We train by optimizing a new loss function obtained from the losses for each task.

To further enhance the performance of our models, we generate three distinct sets of justifications and train models for each set. Ensembles are produced using the output scores for each task to develop more robust labels in the final submission of the run, yielding the other systems submitted for evaluation (*Ens-FT-KE-Mono* and *Ens-FT-KE-Multi*).

**Table 1**
Summary of evaluated systems and their submission number

| System | Submitted Run | System Type |
|---|---|---|
| FS-Ctx | run 1 | Prompting only |
| FT-Base | – | Fine-tuned baseline |
| FT-KE-Mono | – | Fine-tuned |
| FT-KE-Multi | – | Fine-tuned |
| Ens-FT-KE-Mono | run 3 | Ensemble (Fine-tuned) |
| Ens-FT-KE-Multi | run 2 | Ensemble (Fine-tuned) |

## 5. Results

The following section presents the results obtained during the development phase and the final results achieved on the official leaderboard.

### 5.1. Results on dev

We conduct a preliminary evaluation on the development set to validate our proposal and estimate the impact of each component within the proposed methods. Table 2 shows the performance of the generation-based classification and individual fine-tuned models. We observe that the only generative approach does not match the overall performance of the fine-tuned baseline, but still reaches a similar performance in task 1.2. The baseline is outperformed by systems with knowledge expansion, indicating that these systems are indeed learning from the outputs of the generative models. Multitask learning yields similar results to training individual systems, but has the advantage of requiring fewer computational resources, and shared representation benefits task 1.3 slightly.

**Table 2**
ICM-Norm (Hard-Hard evaluation) scores by system and task tested on dev

| System | ICM-Norm Task 1.1 | ICM-Norm Task 1.2 | ICM-Norm Task 1.3 |
|---|---|---|---|
| FS-Ctx | 0.77 | 0.61 | 0.52 |
| FT-Base | 0.80 | 0.61 | 0.56 |
| FT-KE-Mono | 0.81 | 0.64 | 0.58 |
| FT-KE-Multi | 0.81 | 0.64 | 0.59 |

### 5.2. Official leaderboard

System results of the selected systems are presented in Tables 3 and 4. As expected, fine-tuned submitted models outperform their generative counterparts, with the most significant improvement observed in task 1.3. *Ens-FT-KE-Multi* outperforms *Ens-FT-KE-Mono* in tasks 1.1 and 1.3, suggesting that training all tasks might provide insights across tasks that benefit the individual scores, and has the advantage of requiring fewer computational resources. Task 1.2 achieves its best performance with *Ens-FT-KE-Mono*, which may indicate that shared representation is not adding new information to the model.

Results segmented by language are presented in Table 4. Ranking in our system's performance is consistent across languages, but shows differences in metrics, with stronger results in Spanish. This suggests that further experimentation should be conducted in monolingual settings, possibly using techniques such as translation to leverage more data and information.

Our best-performing method achieves second place in all three text-based tasks, showing that combining generative reasoning with fine-tuned classifiers is a promising direction. While the metric performance is encouraging, the results indicate room for improvement and further experimentation.

## 6. Conclusion

This work presents a hybrid approach for sexism identification and characterization in social media, which combines the reasoning capabilities of generative LLMs with fine-tuned transformer-based classification models in a knowledge expansion process.

Key findings include: (1) generative models alone, while competitive, do not surpass fine-tuned classifiers; (2) fine-tuning with knowledge expansion using justifications from generative models of transformer-based models improves performance across all tasks; (3) multitask learning offers computational efficiency while maintaining competitive results and benefiting task 1.3. Our systems achieve competitive results and attain high rankings on the EXIST 2025 leaderboard. The system's performance and internal biases remain to be explored, and future research includes examining other LLMs as components used in the method.

## 7. Ethical Concerns

It is essential to acknowledge that the models developed focus on predicting hard labels, overlooking the granularity of sociodemographic groups' perspectives. In particular, when dealing with violent expressions, prioritizing the voices of victims and vulnerable groups rather than giving equal weight to the views of those who cause harm, helps to surface silenced experiences and attain responsible representation [20]. We also note that LLMs can reproduce various internal biases that might be misleading if not carefully monitored.

**Table 3**
CIMAT-GTO results (Hard-Hard evaluation) across all languages on the official leaderboard. Metrics for the best systems are shown in bold.

| Task | System | Rank | ICM | ICM-Norm | F1 |
|---|---|---|---|---|---|
| 1.1 | FS-Ctx | 34 | 0.5253 | 0.7640 | 0.7639 |
| | Ens-FT-KE-Mono | 3 | 0.6256 | 0.8144 | 0.7968 |
| | Ens-FT-KE-Multi | 2 | **0.6297** | **0.8165** | **0.7996** |
| 1.2 | FS-Ctx | 25 | 0.2069 | 0.5673 | 0.5268 |
| | Ens-FT-KE-Mono | 2 | **0.4678** | **0.6521** | **0.5555** |
| | Ens-FT-KE-Multi | 3 | 0.4392 | 0.6428 | 0.5582 |
| 1.3 | FS-Ctx | 27 | 0.0981 | 0.5228 | 0.5436 |
| | Ens-FT-KE-Mono | 3 | 0.5211 | 0.6210 | 0.6266 |
| | Ens-FT-KE-Multi | 2 | **0.5413** | **0.6257** | **0.6392** |

**Table 4**
CIMAT-GTO results on the leaderboard, segmented by language. Metrics for the best systems in each task are shown in bold. The relative system performance is consistent with the evaluation across all languages.

| Task | Language | System | Rank | ICM-Hard | ICM-Hard Norm | F1 |
|---|---|---|---|---|---|---|
| 1.1 | EN | FS-Ctx | 95 | 0.4757 | 0.7427 | 0.7245 |
| | | Ens-FT-KE-Mono | 6 | 0.5994 | 0.8059 | 0.7744 |
| | | Ens-FT-KE-Multi | 3 | **0.6028** | **0.8076** | **0.7804** |
| 1.2 | | FS-Ctx | 40 | 0.1093 | 0.5378 | 0.4887 |
| | | Ens-FT-KE-Mono | 2 | **0.3287** | **0.6137** | **0.4953** |
| | | Ens-FT-KE-Multi | 7 | 0.2719 | 0.5941 | 0.5002 |
| 1.3 | | FS-Ctx | 42 | 0.0100 | 0.5024 | 0.5145 |
| | | Ens-FT-KE-Mono | 3 | 0.4169 | 0.6022 | 0.6021 |
| | | Ens-FT-KE-Multi | 2 | **0.4190** | **0.6027** | **0.6066** |
| 1.1 | ES | FS-Ctx | 22 | 0.5543 | 0.7772 | 0.7931 |
| | | Ens-FT-KE-Mono | 4 | 0.6351 | 0.8176 | 0.8131 |
| | | Ens-FT-KE-Multi | 3 | **0.6411** | **0.8206** | **0.8136** |
| 1.2 | | FS-Ctx | 25 | 0.2733 | 0.5854 | 0.5542 |
| | | Ens-FT-KE-Mono | 2 | **0.5740** | **0.6793** | **0.6006** |
| | | Ens-FT-KE-Multi | 3 | 0.5628 | 0.6758 | 0.6009 |
| 1.3 | | FS-Ctx | 23 | 0.1652 | 0.5369 | 0.5668 |
| | | Ens-FT-KE-Mono | 3 | 0.6012 | 0.6342 | 0.6449 |
| | | Ens-FT-KE-Multi | 2 | **0.6359** | **0.6420** | **0.6633** |

## Acknowledgments

Villarreal-Haro acknowledges Secretaría de Ciencia, Humanidades, Tecnología e Innovaciones (SECIHTI) for its support provided by the program Becas Nacionales Para Estudios de Posgrados (CVU 1309535). Sanchez-Vega acknowledges SECIHTI for its support through the program "Investigadoras e Investigadores por México" (Project ID.11989, No.1311). The authors thank SECIHTI for the computer resources provided through CIMAT Bajio Super-computing Laboratory (#300832) and Google Cloud Platform for the free cloud credits under its trial program used to access the Gemini API for experimentation in this project.

## Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to perform a grammar and spelling check. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

## References

[1] B. Kitchens, S. L. Johnson, P. Gray, Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption., MIS quarterly 44 (2020).

[2] D. L. Rhode, The subtle side of sexism, Colum. J. Gender & L. 16 (2007) 613.

[3] J. K. Swim, L. L. Hyers, L. L. Cohen, M. J. Ferguson, Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies, Journal of Social issues 57 (2001) 31–53.

[4] J. Fox, C. Cruz, J. Y. Lee, Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media, Computers in human behavior 52 (2015) 436–442.

[5] S. Dunn, Technology-facilitated gender-based violence: An overview (2020).

[6] L. I. Molnar, "i didn't have the language": Young people learning to challenge gender-based violence through consumption of social media, Youth 2 (2022) 318–338.

[7] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and TikTok videos, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction., Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[8] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and TikTok videos (extended overview), in: CLEF 2025 Working Notes, 2025.

[9] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[10] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[11] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023-learning with disagreement for sexism identification and characterization (extended overview)., CLEF (Working Notes) (2023) 813–854.

[12] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[13] Z. Zhang, Z. Zeng, Y. Lin, H. Wang, D. Ye, C. Xiao, X. Han, Z. Liu, P. Li, M. Sun, et al., Plug-and-play knowledge injection for pre-trained language models, arXiv preprint arXiv:2305.17691 (2023).

[14] Z. Song, B. Yan, Y. Liu, M. Fang, M. Li, R. Yan, X. Chen, Injecting domain-specific knowledge into large language models: a comprehensive survey, arXiv preprint arXiv:2502.10708 (2025).

[15] A. Nirmal, A. Bhattacharjee, P. Sheth, H. Liu, Towards interpretable hate speech detection using large language model-extracted rationales, in: Y.-L. Chung, Z. Talat, D. Nozza, F. M. Plaza-del Arco, P. Röttger, A. Mostafazadeh Davani, A. Calabrese (Eds.), Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 223–233. URL: https://aclanthology.org/2024.woah-1.17/. doi:10.18653/v1/2024.woah-1.17.

[16] K. Villarreal-Haro, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Stacked reflective reasoning in large neural language models, Working Notes of CLEF (2024).

[17] K. Villarreal-Haro, G. Segura-Gómez, J. Tavarez-Rodríguez, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Leveraging reasoning of auto-revealed insights via knowledge injection and evolutionary prompting for sexism analysis, Working Notes of CLEF (2025).

[18] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, et al.,

The prompt report: A systematic survey of prompting techniques, arXiv preprint arXiv:2406.06608 5 (2024).

[19] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022), European Language Resources Association, Marseille, France, 2022, pp. 258–266.

[20] J. Butler, Precarious life: The powers of mourning and violence, verso, 2004.