

Transfer Learning from Multilingual DeBERTa for Sexism Identification

Hoang Thang Ta^{1,2}, Abu Bakar Siddiquir Rahman^{1,3,*}, Lotfollah Najjar³ and Alexander Gelbukh¹

¹*Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico*

²*Dalat University, Lam Dong, Vietnam*

³*University of Nebraska Omaha, Omaha, Nebraska, USA*

Abstract

In this paper, we address the Task 1 and Task 2 of the EXIST 2022 in detecting sexism in a broad sense, from ideological inequality, sexual violence, misogyny to other expressions that involve implicit sexist behaviours in social networks. We apply transfer learning from a pre-trained multilingual DeBERTa (mDeBERTa) model and its zero classification to gain a better performance than BERT-based approaches. Lastly, we combine all 3 methods: mDeBERTa, zero classification, and BERT for majority vote. For Task 1, mDeBERTa is the best method with an accuracy of 76.09% and F1 of 76.08%. Meanwhile, an accuracy of 66.26% and F1 of 47.06% are the best results in Task2, when using majority vote. Our main contribution is to use DeBERTa and zero classification with designing only one classifier in sexism identification.

Keywords

Sexism Identification, DeBERTa, Text Classification, Offensive Language, EXIST 2022, IberLEF

1. Introduction

In the field of sentiment analysis, sexism identification plays an essential role in observing how modern society discriminates against people based on gender. The expeditious nature of human psychology about the interactions between people causes a widespread use of all online social media platforms. As online social platforms have the accessibility to express people's opinion without meeting in person and to some extent, it is possible to hide the real identity, it experienced abusive language from dissimilar aspects of society. Among these, sexism attitudes towards women are immensely menacing. Specifically, woman can recognize sexually ambiguous comments more likely as a sexual harassment [1, 2]. On social media posts perspective, Chowdhury et. al. recollected sexual harassment posts from Twitter for specific user to identify which users were related on the posts of sexual harassment [3]. Sexism is responsible for women's psychological disorder by reducing their comfort zone to make

IberLEF 2022, September 2022, A Coruña, Spain.

*Corresponding author.

✉ tahoangthang@gmail.com (H. T. Ta); abubakarsiddiquirra@unomaha.edu (A. B. S. Rahman);

lnajjar@unomaha.edu (L. Najjar); gelbukh@cic.ipn.mx (A. Gelbukh)

🆔 0000-0003-0321-5106 (H. T. Ta); 0000-0002-8581-0891 (A. B. S. Rahman); 0000-0003-3960-4189 (L. Najjar);

0000-0001-7845-9039 (A. Gelbukh)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

comments publicly, increasing their anger and depression from bullying posts, and hence that effects them to think themselves not considered as a self-esteem human in the society [4]. Sexism behaviour noticed in a plummeting manner to every aspects of society. In the United States, one of ten adult females are victims of sexism, considered as harassment that conduct a counterbalance impact on the society [5]. To mitigate these problems, there needs to have a solution to minimize the spreading of the harassment behaviors from social media. Artificial intelligence and deep learning approaches draw a beneficial trade to identify and classify the sexism texts from social media by solving these problems.

Similar to the previous challenge, EXIST 2022 (<http://nlp.uned.es/exist2022/>) requires participants to identify sexism in English and Spanish tweets, divided into 2 tasks [6]. Task 1 will detect whether a tweet contains sexism or not, in the form of a binary classification. In contrast, Task 2 is a multiclass classification which continues to detect which sexism categories a given tweet belong to if it contains sexism. There are 5 sexism categories in total, they are ideological-inequality, stereotyping-dominance, objectification, sexual-violence, misogyny-non-sexual-violence.

In this paper, we apply transfer learning from multilingual pre-trained models to design one classifier which can solve both tasks of sexism identification at the same time. We focus our work on DeBERTa and its zero classification for comparing the model performance with BERT. We also combine these three methods (DeBERTa, zero classification, and BERT) in majority vote to gain the best results. Besides this section, Section 2 introduces related works and popular methods to detect sexism in short texts. Section 3 and Section 4 describe the tasks in detail and analyze the datasets obtained from organizers. Our methodology used to train the models and produce the results is presented in Section 5. Lastly, we report our experiments, as well as withdraw conclusions and declare future works, in Sections 6 and Section 7 correspondingly.

2. Related Works

Basically, sexism identification belongs to abusive language, a sub-field of sentiment analysis. It also has a close relation with several types of abusive language, such as racism, hate speech, personal attacks and others [7]. We consider sexism identification as a problem of text classification, where the classifiers will recognize which predefined labels that a given text or tweet belong to. Sexism detection can detect efficiently with methods used in natural language processing (NLP). Recently, machine learning and deep learning techniques were used for multi-label classification.

Rodriguez et. al. proposed the task of sexism identification, developed MeTwo dataset, and applied machine learning techniques (traditional and deep learning methods) to detect types of sexist behaviours [8]. Since then, they opened a challenge of the sexism identification in last year. In the previous challenge, there was a wide diversity of approaches/methods used in the participation, from transformer models such as BERT, BETO [9] (a version of Spanish BERT), RoBERTa, XLM-R [10] to traditional machine learning methods such as Support Vector Machines (SVM), Random Forest (RF), or Logistic Regression (LR), and even other deep learning methods, such (i.e. Long short-term memory networks - LSTM) and with the fastText library [11]. Several top teams apply multi-task learning [9] and external datasets [10] to

strengthen or generalize the model, hence achieve higher results. Aside from challenges, some other research works done to detect sexism from text in social media platform. Samory et. al. used logistic regression, CNN and BERT fine tuned to detect sexism from twitter dataset with the baseline of gender specific words whether it exists in a sentence or not and toxicity based text that use toxicity ratings with Jigsaw's perspective API [12].

There are some works, also about sexism, related to the challenge. To boost the model performance, Sharifirad et. al. applied text augmentation and text generation on tweets from ConceptNet and Wikidata [13]. Their research detected sexism in tweets with 4 types: indirect harassment, information threat, sexual harassment and physical harassment. Jha and Mamidi used Support Vector Machines (SVM), sequence-to-sequence models and FastText classifier to identify sexual tweets in 3 categories: hostile, benevolent and others [14]. A dataset was created from Everyday Sexism Project website that consists of posts of survivors and observers accounts. The dataset labelled by 14 categories based on some policy on social scientist with ten annotators who has knowledge about gender and/or sexuality. The 14 categories of the dataset: role stereotyping, attribute stereotyping, body-shaming, hyper-sexualisation, internalized sexism, hostile work environment, denial or trivialization of sexual misconduct, threats, sexual assault, sexual harassment excluding assault, moral policy and victim blaming, slut shaming, motherhood and menstruation related discrimination and other [15]. In a multi-label classification view, Parikh et. al. provided a new dataset with 23 categories over 13023 accounts of sexism [16]. Karlekar et. al. was collected a dataset from a publicly available online forum named by SafeCity based on Sexual harassment stories. The dataset contained 9892 stories with 13 forms of harassment type, however, only three types, grouping, oggling and toxic commenting were used to automatically categorizing and analyzing harassment type by using the combination of CNN and RNN models for the the multi label classification tasks [17].

3. Task Description

Participants are required to classify "tweets" in both languages, English and Spanish. Task 1 is a binary classification, which one needs to build a system to detect whether a given tweet containing sexist expressions or not.

With detected tweets to category sexist in Task 1, participants must continue to detect which sexism categories they belong in Task 2, with a theme of multiclass classification. There are 5 sexist categories:

- ideological-inequality: This sexist type denied the role of equality between men and women, defame the feminism, or consider men as the center of gender oppression. For example, *"I hate men that be in women's business. Acting like a bitch. Shut up."*
- stereotyping-dominance: This type contains wrong opinions about woman, considering they are more appropriate in the certain roles (housewife, family caregiving, wife, mother, tender, etc). Furthermore, it declares that the woman can not work as the men in some tasks (construction, driving, technical, politics, etc). For example, *"Guy who thinks women shouldn't be allowed to serve in Star fleet"*.
- objectification: The text considers women as objects aside from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order

to fulfill traditional gender roles (compliance with beauty standards, hypersexualization of female attributes, women’s bodies with wants of men, etc.). For example, *”You look like a prostitute whore Nancy.”*

- sexual-violence: The text contains sexual aspects such as sexual suggestions, sexual assault, harrassment or even rape towards woman. For example, *”Yes I am European and you were born one of the lowest races you Dravidian slut.”*
- misogyny-non-sexual-violence: This type contains tweets related non-sexual violence and animosity against woman. For example, *”I sentence her to be gangbanged to death by the daily stormer editorial team.”*

4. Dataset Analysis

In the challenge 2022, the training set will be the training and test sets of 2021. After merging these two sets, there are 11,345 tweets in total, with 5701 of English and 5644 of Spanish. Then, the organizers released a test set with 1,058 tweets. The distribution of tweets by categories in the training set is shown in Figure 1.

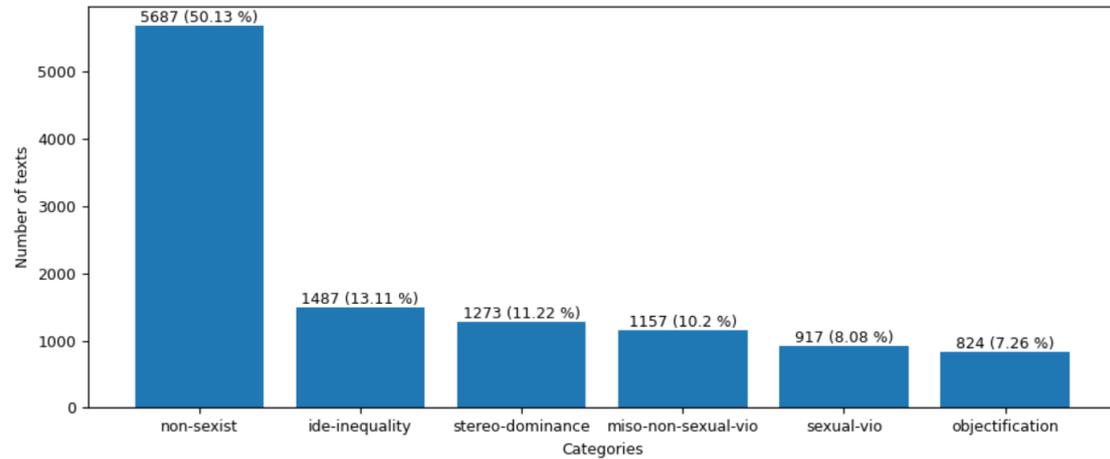


Figure 1: The distribution of tweets by categories.

For non-sexist category, there are 5687 tweets, accounted for 50.13% while sexist categories take 49.87% with 5658 tweets. If count only for sexist categories, there are 1487, 1273, 1157, 917, and 824 tweets in corresponding to ideological-inequality, stereotyping-dominance, objectification, sexual-violence, and misogyny-non-sexual-violence categories. Clearly see that, the training dataset now is imbalanced, however we do not apply any oversampling or undersampling methods in this paper to see how well mDeBERTa can perform on the sexism identification.

There are a few tweets containing a large number of tokens, up to over 700, while most of the tweets have from 1 to 80 tokens as in Figure 2. Even so, we still set `MAX_LEN=400` in the training process to support long tweets. If a tweet has more than `MAX_LEN` value, the classifier will take only first 400 tokens.

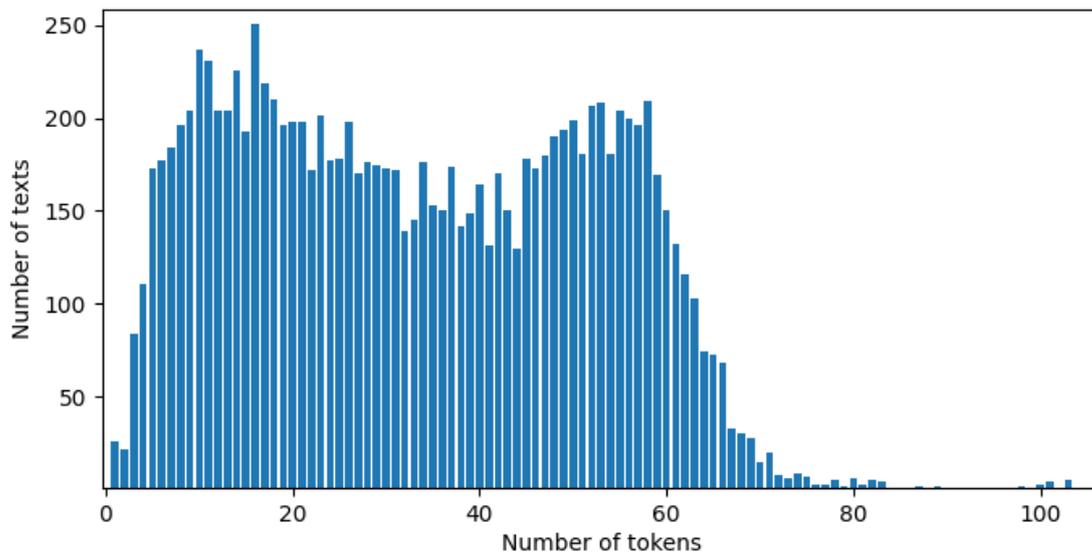


Figure 2: The distribution of tweets by the first 100 number of tokens.

5. Methodology

5.1. Preprocessing

Before putting tweets into the training process, a few steps of data preprocessing are used. Our purpose is not to distort the original meaning of tweets from impacting on their semantic components, we thus only a light preprocessing with a list of steps:

- Split urls which are sticky in text by blanks. For a long url, we split its into a sequence of words. We do not remove urls because we realize some urls keeping the helpful information to identify sexism in tweets. For example, url `http://casoaislado.com/refugiado-...-intento-violacion/` can extract to a list of words as "refugiado afgano viola una nina 13 anos tras salir la carcel intento violacion".
- Normalize hashtags by removing #. For example, hashtags `#SpeakFreely`, `#MAGA`, `#GabFam` will be normalized as `SpeakFreely`, `MAGA`, and `GabFam`.
- Normalize `@[user]` to `@username` for English tweets and `@usuario` for Spanish tweets.
- Apply 2 language packages (`en_core_web_md` and `es_core_news_md`) of `spaCy v2.3.2` (<https://spacy.io/>) to analyze tweets into tokens, also remove redundant spaces.

5.2. Classifier design

From the original dataset, the distribution of tweets is reorganized into 6 categories: non-sexist, ideological-inequality, stereotyping-dominance, objectification, sexual violence, and misogyny-non-sexual-violence. By this way, we can only train a model instead of separating into two models for the tasks. For Task 1, the model will predict category non-sexist

Table 1

The result table of Task 1 by runs from EXIST2022 organizers.

#	English	Spanish
1	sexist	sexista
2	non-sexist	no sexista
3	inequality	desigualdad
4	dominance	dominio
5	objectification	objetivación
6	violence	violencia
7	misogyny	misoginia

and the remain categories as sexist. For Task 2, the model only predict exactly which sexism categories that a given tweet belong to.

5.3. Zero-shot classification

Zero-shot classification is a great way for the problem of text classification without requiring to train models or fine-tune pre-trained models by your own needs. Here, we can freely define the labels but this makes the model performance depending on how to choose the correct labels. We create a classifier from pipeline zero-shot-classification from transformer package on pre-trained model `MoritzLaurer/mDeBERTa-v3-base-mnli-xnli`. Given a tweet t and a list of labels $L=\{l_1, l_2, \dots, l_n\}$, the model output will be a list of probabilities:

$$\sum_i^n P(t, l_i) = P(t, l_1) + P(t, l_2) + \dots + P(t, l_n) = 1 \quad (1)$$

where n is the size of the label list. If allow multiple true labels, these probabilities will be independent, having a value in range $[0, 1]$.

To treat tweets in English and Spanish fairly, we translate category labels from English to Spanish and apply them for the classification. In some experiments, we realize that zero-shot classification can not work well on labels with noun phrases. We thus truncate the original labels, pick single important words that can differentiate from other labels as in Table 1. In this paper, our runs on zero-shot classification will be not estimated by organizers due to the submission’s limitation. However, it can contribute to obtain the best performance through out majority vote.

5.4. Multilingual pre-trained DeBERTa models

DeBERTa (decoding-enhanced BERT with disentangled attention) is a model architecture which offers disentangled attention and enhanced mask decoder to improve the model performance, compared to BERT and RoBERTa models. It is also a large scale pre-trained language model containing 1.5 parameters, and a research project funded by Microsoft (<https://www.microsoft.com/en-us/research/project/deberta/>).

In the attention mechanism, BERT uses only one vector from the sum of word embedding and position embedding for a given word. Otherwise, two vectors are used to represent a word in DeBERTa, one for word embedding and the other for position embedding. Then, we can calculate the attention weights among words by disentangled matrices based on their contents and relative positions.

Similar to BERT, DeBERTa is also pre-trained on masked language modeling (MLM), which the model will take the content and position of context words for the training. By this way, we can see that the disentangled attention mechanism allows to have a better prediction in some cases. For example, the words `store` and `mall` are masked for the prediction in the sentence "a new store opened beside the new mall". Normally, these two words play a similar role in the local contexts. However, in the aspect of syntactic, `store` is a subject while `mall` is not [18]. The newest version of DeBERTa, DeBERTaV3 apply the replacements of mask language modeling (MLM) with replaced token detection (RTD), to improve the effectiveness in pre-training tasks [19].

In this challenge, we try a new approach, multilingual DeBERTa (mDeBERTa) for the problem of sexism identification. For more convenience, we search for mDeBERTa models which are available on HuggingFace in a similar domain of the challenge and in Spanish (<https://huggingface.co/models?language=es&sort=downloads&search=deberta>). Unfortunately, we can see the only one pre-trained model, `MoritzLaurer/mDeBERTa-v3-base-mnli-xnli`. Therefore, we apply it to the experiment, described in detail in the next section. We also use multilingual BERT models with comparison purposes.

6. Experiments

Since the dataset contains tweets in English and Spanish, we thus choose to fine-tune two multilingual pre-trained models on tasks of sexism identification.

- `nlptown/bert-base-multilingual-uncased-sentiment`: This model is finetuned from a `bert-base-multilingual-uncased` model for sentiment analysis on product reviews in six languages: English, Dutch, German, French, Spanish and Italian. It predicts the sentiment of the review as a number of stars (between 1 and 5). We name this model in short is `bert-base-multilingual`.
- `MoritzLaurer/mDeBERTa-v3-base-mnli-xnli`: We apply this model to multilingual zero-shot classification and to fine-tune on special tasks related to sentiment analysis. The model works on 100 languages in the field of natural language inference (NLI). It was initially trained by Microsoft on the CC100 multilingual dataset, then fine-tuned on the XNLI dataset (hypothesis-premise pairs from 15 languages) and the English MNLI dataset. We name this model in short is `mDeBERTa-v3-base`.

The model configuration uses `MAX_LEN = 400`, `RANDOM_SEED = 42`, `DROP_OUT = 0.1`. Due to the limitation of computer resources on our server, we must use `BATCH_SIZE = 4`. In the training, we apply linear regression on embeddings to fit `hidden_size` (768 as default) to `n_classes` (6, the number of categories). The optimizer is AdamW with `learning_rate=1e-5` and `CrossEntropyLoss()` is used to calculate the average loss.

Table 2

The result table of our runs from EXIST2022 organizers.

#	Model/Method	Submission file	Task	Acc	F1	Rank
1	majority vote	task1_ThangCIC_7.tsv	Task 1	0.7609	0.7608	21/47
2	mDeBERTa-v3-base	task1_ThangCIC_3.tsv	Task 1	0.7609	0.7600	22/47
3	bert-base-multilingual	task1_ThangCIC_1.tsv	Task 1	0.7580	0.7553	24/47
4	BASELINE	–	Task 1	0.6928	0.6859	39/47
5	Majority Class	–	Task 1	0.5444	0.3525	44/47
6	zero-shot classification	task1_ThangCIC_5.tsv	Task 1	–	–	–
#	Model/Method	Submission file	Task	Acc	F1	Rank
1	majority vote	task2_ThangCIC_8.tsv	Task 2	0.6626	0.4706	10/31
2	mDeBERTa-v3-base	task2_ThangCIC_4.tsv	Task 2	0.6551	0.4612	15/31
3	bert-base-multilingual	task2_ThangCIC_2.tsv	Task 2	0.6626	0.4562	16/31
4	BASELINE	–	Task 2	0.5784	0.3420	26/31
5	Majority Class	–	Task 2	0.5539	0.1018	30/31
6	zero-shot classification	task2_ThangCIC_6.tsv	Task 2	–	–	–

By doing some experiments on the training and validation sets of last year challenge, we conclude that the splitting did not give a better model performance. Therefore, we train models on full dataset to help it learn as much as possible the available data, although we can not see or prevent the overfitting if happen. The models was trained on 30 epochs and we save on disk the best model whenever gaining the highest accuracy values.

Our results are shown in Table 2 for different configurations in the training process. The organizers provide BASELINE and Majority Class methods, which underperform significantly all of our runs. As mentioned in Section 5.3, we are not able to get the results from zero-shot classification runs because each team can only submit three runs for each task. In both tasks, DeBERTa is slightly better than BERT about 0.5% in term of F1. When combining all three methods (DeBERTa, BERT, and zeroshot-classification) in majority vote, the model can earn the best performance in both tasks. Follow that, the best results of Task 1 are the accuracy of 76.09% and F1 of 76.08%, while Task 2 has the accuracy of 66.26% and F1 of 47.06%.

7. Conclusion

In this paper, we present an approach of transfer learning from multilingual transformer models, especially DeBERTa to solve in the problem of sexism identification. The input data contains some minor steps of preprocessing before putting the training mode on 2 pre-trained models, bert-base-multilingual and mDeBERTa-v3-base. In general, DeBERTa is slightly better than BERT on both tasks. Furthermore, when combining these two with zero-shot classification in majority vote, the model obtains the best performance in Task 2 and a very close result to DeBERTa in Task 1. In Task 1, our best accuracy is 76.09% and the best F1 is 76.00%, while 66.26% and 47.06% are the best values of accuracy and F1 on Task 2. Our main contribution is to bring a new transformer model (DeBERTa) for the training and to classify tweets by a multilingual approach on both tasks at the same time.

Although, our model performance is ranked in middle of the result table, the gap between us and top teams is not large. The results are also hinted that sexism identification is a difficult task which popular models alone can not solve the problem effectively. Any standard model can outperformed the majority vote approach but to be the best is another story. We thus must treat tweets with a different view instead of a sequence of words and work on other approaches such as multi-task learning and weak label supervision to design better classifiers which can reach closer to human evaluation. That is our work in the future.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

- [1] J. A. Hehman, C. A. Salmon, A. Pulford, E. Ramirez, P. K. Jonason, Who perceives sexual harassment? Sex differences and the impact of mate value, sex of perpetrator, and sex of target, *Personality and Individual Differences* 185 (2022) 111288.
- [2] J. K. Swim, R. Mallett, C. Stangor, Understanding subtle sexism: Detection and use of sexist language, *Sex roles* 51 (2004) 117–128.
- [3] A. G. Chowdhury, R. Sawhney, R. R. Shah, D. Mahata, YouToo? detection of personal recollections of sexual harassment on social media., in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2627–2537.
- [4] S. H. Berg, Everyday sexism and posttraumatic stress disorder in women: A correlational study, *Violence Against Women* 12 (2006) 970–988.
- [5] 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.
- [6] Francisco Rodríguez-Sánchez, Jorge Carrillo-de-Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, Paolo Rosso, Overview of exist 2022: sexism identification in social networks, *Procesamiento del Lenguaje Natural* 69 (2022).
- [7] M. Karan, J. Šnajder, Cross-domain detection of abusive language online, in: *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 2018, pp. 132–137.
- [8] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, *IEEE Access* 8 (2020) 219563–219576.
- [9] F. M. Plaza-del Arco, M. D. Molina-González, L. Alfonso, Sexism Identification in Social Networks using a Multi-Task Learning System (2021).
- [10] S. Mina, B. Jaqueline, L. Daria, S. Djordje, K. Armin, H. Manuel, B. Johannes, S. Sven,

- S. Alexander, Z. Matthias, Automatic sexism detection with multilingual transformer models, arXiv preprint arXiv:2106.04908 (2021).
- [11] 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.
- [12] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, Call me sexist, but...: Revisiting sexism detection using psychological scales and adversarial samples, in: *Intl AAAI Conf. Web and Social Media*, 2021, pp. 573–584.
- [13] S. Sharifirad, B. Jafarpour, S. Matwin, Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs, in: *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 2018, pp. 107–114.
- [14] A. Jha, R. Mamidi, When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data, in: *Proceedings of the second workshop on NLP and computational social science*, 2017, pp. 7–16.
- [15] P. Pulkit, Abburi, Harika, G. M. Niyati, Chhaya, V. Varma, Categorizing Sexism and Misogyny through Neural Approaches, in: *Categorizing Sexism and Misogyny through Neural Approaches. ACM Transactions on the Web (TWEB)*, 2021, pp. 1–31.
- [16] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, V. Varma, Multi-label categorization of accounts of sexism using a neural framework, arXiv preprint arXiv:1910.04602 (2019).
- [17] S. karlekar, B. Mohit, SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories, 2018, pp. 2805–2811.
- [18] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).
- [19] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).