Emotion Signals for Sexist and Offensive Language Detection: A Multi-task Learning Approach

Conference Paper · September 2023

CITATIONS
0

READS
55

3 authors, including:

Alexandra Ciobotaru
University of Bucharest
6 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Diana Constantina Hoefels
University of Tuebingen
3 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Emotion Signals for Sexist and Offensive Language Detection: A Multi-task Learning Approach

Alexandra Ciobotaru,[1] Diana Constantina Höfels,[2] Ștefan Daniel Dumitrescu[3]

[1]University of Bucharest and DRUID AI, alexandra.ciobotaru@unibuc.ro
[2]University of Tübingen, diana-constantina.hoefels@uni-tuebingen.de
[3]Adobe, sdumitre@adobe.com

Abstract

Identification and classification of sexist and offensive content in social media posts present a great deal of complexity and relevance. Detecting and identifying this type of language is more difficult due to the presence of multiple forms of sexist and offensive language. We employ a multi-task learning framework to link emotion detection to sexist and offensive language classification, allowing the two tasks to complement each other. The results of our study demonstrate that the use of emotion signals aids the performance of sexist and offensive language detection - we achieve an F1-score of 87.98% by fine-tuning the Romanian BERT, which becomes the state of the art for sexism and offensive detection in Romanian short texts.

Keywords— MTL, sexist and offensive language detection, emotion detection, BERT, Romanian corpora, low-resourced languages

1. Introduction

Social media platforms have profoundly altered the manner in which we communicate, and these shifts have given rise to egregious practices online, such as the use of offensive or sexist language. A sexist act is a discrimination against a person on the basis of their gender. A wide range of sexist language forms, both overt and covert, affect mostly women and girls across many areas of life, including the workplace (Verniers and Vala, 2018), politics, society, household responsibilities, and even Natural Language Processing (NLP) models (Sun et al., 2019). Despite the lack of a universally accepted definition of offensive language, it is commonly associated with cursing, profanity, blasphemy, epithets, obscenity, and insults (Jay, 1992). Thus, automatically detecting discriminatory language can assist in analyzing it so that preventative measures can be taken. In terms of gender representation in language, most feminist language activists support the change of language as a means of achieving better gender equality (Pauwels, 2003). The use of these systems can be useful in the development, design, and dissemination of policies related to equality, as well as in contributing to social change in a positive direction.

Identifying subtle forms of sexist and offensive language can be quite challenging. At present, the majority of research focuses on each of these tasks individually. Despite their compelling results, these approaches are limited to modeling only the linguistic aspects of discriminatory and offensive language, without taking into account an important aspect, such as emotions. Emotions are highly prevalent in language and thought. Using information obtained from people's emotional states when expressing themselves could support and improve the development of natural language applications. For more complex semantic tasks, such as detecting sexist and offensive language, a unified system may be necessary. The majority of systems, however, do not possess certain features that may help facilitate the development of such a system, as discussed above. The concept of Multi-Task Learning (MTL) is based on human learning activities in which individuals apply their knowledge from auxiliary tasks to assist them in the learning of a new task. It is an approach of inductive transfer in which the domain information contained in the training signals of related tasks is used as an inductive bias to facilitate generalization (Caruana, 1997). At its infancy, MTL is motivated primarily by the goal of alleviating the problem of data sparsity, and by aggregating the labeled data in all tasks, MTL achieves more accurate learning for each task, being therefore useful in reusing existing knowledge and reducing the cost of manual labeling. Lastly, deep MTL models perform better than single-task models (Zhang and Yang, 2022). Our study hypothesizes that adding emotion information to the mix can help in detecting sexism and offensive language in Romanian tweets in a more efficient and effective manner.

Furthermore, most of the sexist and offensive language detection systems are developed for well-resourced languages. Therefore, a key objective of this paper is the development of language technologies in the midst of a scarcity of digital language resources and tools for Romanian language. Romanian is a Romance language spoken by approximately 24 to 26 million people as a native language, while about 4 million speak it as a secondary language.[1]

To the best of our knowledge, this is the first study to employ emotion analysis in the detection of sexist and offensive language in a less-resourced language, such as Romanian.

2. Related Works

Over the past few years, there have been numerous academic events and shared tasks related to the identification of sexist and offensive language. For low-resource languages, however, the detection of offensive language has received relatively little attention in NLP.

The EXIST (Sexism Identification in Social Networks) competition at IberLEF (Rodríguez-Sánchez et al., 2022) was the first collaborative effort aimed at detecting sexism in a broad sense, from outright misogyny to more subtle expressions of sexism. Within the same shared task, del Arco et al. 2021 test the performance of a multi-task learning approach that incorporates sentiment analysis and offensive language detection to identify sexism. In another recent study, del Arco et al. (2022) investigated more linguistic phenomena than sentiment analysis in their research for sexism detection. Using multi-task methods they incorporate emotions, sarcasm, insults, constructiveness, and targets into the learning process.

[1]https://www.britannica.com/topic/Romanian-language

Sharifirad et al. (2019) examine the users' mood when writing sexist tweets. They use the SemEval-2018 task1: Affect in tweets dataset (Mohammad et al., 2018), and examine the types and intensities of emotions associated with categories of sexual harassment. According to their findings, indirect harassment, also known as benevolent harassment, has a mild intensity. On the other hand, hostile sexism is associated with very high levels of disgust, anger, sadness, and even joy. The tweets also demonstrate that users enjoy sending sexist messages to women.

In modelling the linguistic properties of abusive language, Rajamanickam et al. (2020) consider the emotional state of the users and how this may affect their language. Using a multi-task learning framework, they present a joint model of emotion detection and abusive language detection. According to their results, incorporating affective features increases abuse detection performance across datasets significantly.

Using three auxiliary tasks that were automatically created through unsupervised learning from a set of unlabeled and weakly labelled accounts, Abburi et al. (2020) explored neural multitask learning and investigated 23-class fine-grained classifications of accounts of sexism.

3. Corpora

For elaborating the multi-task architecture we used two datasets, CoRoSeOf (Hoefels et al., 2022), and REDv2 (Ciobotaru et al., 2022).

CoRoSeOf is a large corpus of Romanian social media manually annotated for sexist and offensive language. There are covert and overt forms of sexist language included in the corpus, which have been classified into direct, descriptive, and reported statements. It consists of 39245 tweets which have been annotated with the following labels: *sexist direct*, *sexist descriptive*, sexist reporting, *non sexist offensive* and *non sexist*. It is important to note that approximately 80% of this corpus is skewed towards the non sexist class.

To compliment the sexist and offensive language classification by using a multi-task approach, we construct the auxiliary task using the RED dataset. REDv2 is a Romanian emotion detection dataset containing 5449 tweets annotated in a multi-label fashion, with the following emotions: *anger*, *fear*, *joy*, *sadness*, *surprise*, *trust* and *neutral*.

4. System overview

As a preliminary step towards predicting the emotions in each CoRoSeOf text, we developed an emotion detection model by fine-tuning the cased version of the Romanian BERT (Dumitrescu et al., 2020) from Huggingface[2] on the task of classifying emotion labels of REDv2 tweets.[3]

We have created a model that loaded the weights from the bert-base-cased with a linear layer on top, using Transformer's class AutoModelForSequenceClassification, with its *from_pretrained* method, and training was conducted using HuggingFace's Trainer API. The model was trained for five epochs with a batch size of eight and a learning rate of 2e-5. The loss function used was BCEWithLogitsLoss, which combines the sigmoid layer with the Binary Cross Entropy loss in a single class. As a result of this model, we were able to obtain an F1-score of 0.71 on REDv2.

Figure 1 illustrates the distribution of emotion labels in the five CoRoSeOf classes using UpSet plots (Lex et al., 2014). These plots are used to visualize intersections between more than three sets, in a matrix. The horizontal bars in the left part of the matrix represent the total number of texts which contain the specified emotion in each row. Using the vertical bar above the matrix, each unconnected dot on the matrix shows the count of texts containing one unique emotion, while the connected dots indicate the number of texts sharing more than one emotion.

In applying an upset plot to CoRoSeOF texts after emotion prediction, the following insights can be derived: *non sexist* tweets are mainly neutral while in *non sexist offensive* tweets, anger is the predominant emotion expressed, followed by neutral, sadness, and a mixture of sadness and anger. The majority of *sexist direct* tweets express joy, anger, and neutral feelings, while a few express trust, and the combination of trust and joy. *Sexist descriptive* tweets are mainly neutral, followed by a relatively high number that express anger, and a few express trust, sadness, joy, as well as combinations of trust and neutral, sadness and anger, and trust and joy. There is a predominance of anger in *sexist reporting* texts, with neutral representing the second most prevalent emotion.

The first column in each upset plot represents the amount of texts in the specified category which have not received an emotion label by the emotion detection model, because the probability of detection for each of the seven emotion labels was lower than 50%.

Figure 2 outlines the three types of model architectures that were considered. In order to develop a baseline, we start with the simplest architectural model; then, we add extra data to the model in the form of precomputed emotion probabilities; and we compare the results of that model with those of a multi-task model trained on the CoRoSeOf and REDv2 datasets jointly.

Finally, training was standardized, i.e., all models use a 0.1 dropout, the same learning rate and learning schedule, and the same early stopping criterion and patience (including the multi-task model where we early stopping considers only the validation set of CoRoSeOf.

4.1. Data Preprocessing

In order to ensure accuracy and minimal bias in the data, we preprocessed it using the following steps: first, we removed all usernames from the CoRoSeOf dataset. The nature of Twitter responses prompted us to take this action; since they are branched off from the original tweet in a tree-like fashion (Ryosuke Nishi (2016)), it was paramount to avoid our models to be biased towards certain users. Secondly, we deleted from CoRoSeOf 49 texts that had a majority vote ground truth of 'Cannot decide' and 1457 texts with 'Non agreement'. Thirdly, we replaced names with "person" using `roner` python library (Dumitrescu and Avram (2019)), emails with "email" using regex, and also we eliminated telephone numbers, as these do not bring valuable information in the machine learning process and it was important to align the CoRoSeOf preprocessing with the REDv2 preprocessing. Lastly, we split the data in an 80/10/10 fashion, making sure each CoRoSeOF label has an equal distribution for training, testing and validation.

4.2. Baseline Model

The baseline model is a BERT transformer from which the pooled output is forwarded to a dense layer representing the output classes. In the training process, there is a standard dropout of 0.1 before the last layer, in which the cross-entropy loss is computed. This is the most basic model that can be used with transformers; although simple, it provides a strong baseline.

[2] `https://huggingface.co/dumitrescustefan/bert-base-romanian-cased-v1`

[3] `https://huggingface.co/datasets/Alegzandra/REDv2`

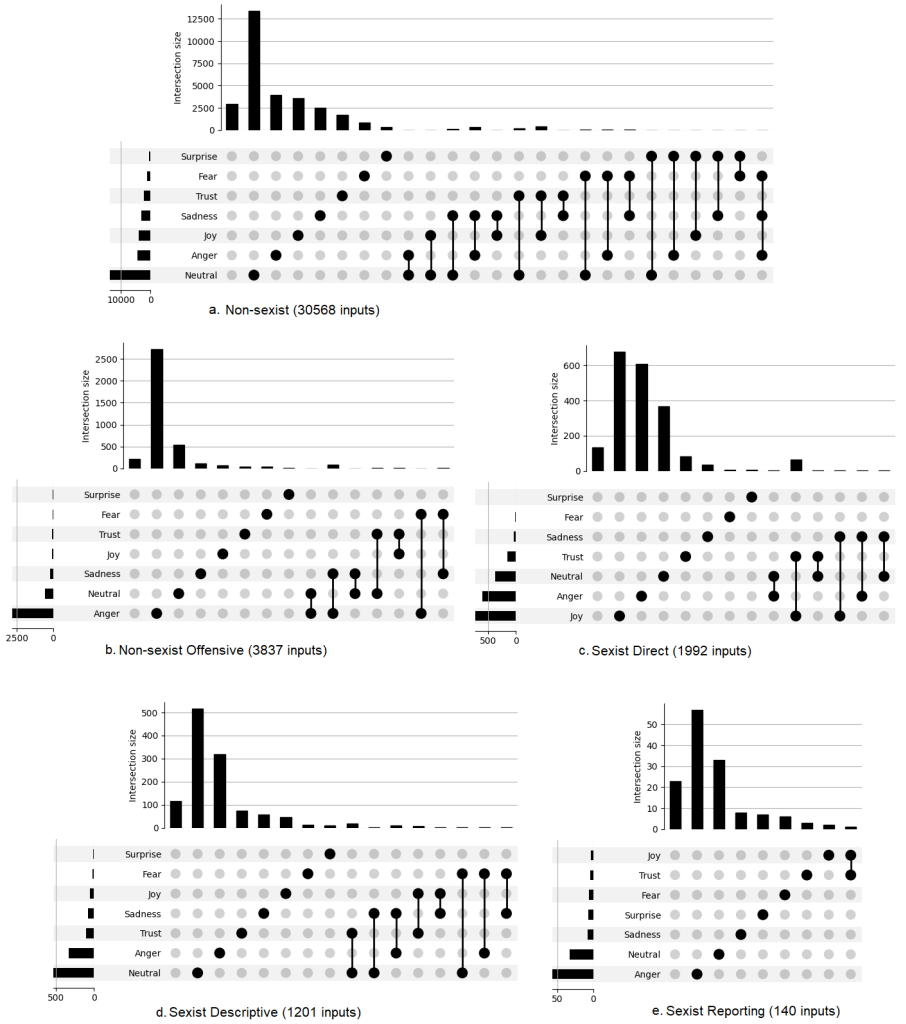

Figure 1: Multi-Label Emotion Distribution in CoRoSeOf Classes: a. non sexist, b. non sexist Offensive, c. Sexist Direct, d. Sexist Descriptive, e. Sexist Reporting.

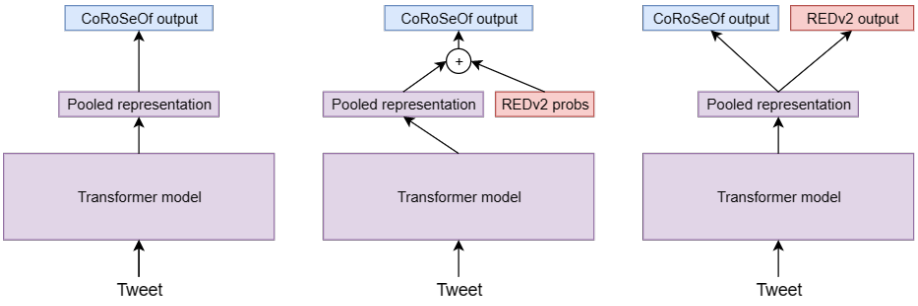

Figure 2: The three architectures tested: baseline (A), feature engineered model where we add the pre-computed REDv2 emotion probabilities for each tweet (B), and the multi-task model with distinct heads for each dataset (C)

4.3. Feature-engineered Model

The feature-engineered model benefits from the results of the existing emotion detection model trained on REDv2, by making use of the emotion prediction for each tweet, in addition to the tweet's text itself.

Architecturally, the model is identical to the baseline, but after pooling all the outputs from the transformer into a single-dimensional 768 vector (as with bert-base), we concatenate seven more values to it, representing the floating point probability predictions of the seven emotions of REDv2. To summarize, we add these seven pre-computed values to the input of the class-output layer.

4.4. Multi-task Model

Our multi-task model architecture follows the standard paradigm where we have a single model that encodes tweets, but different output heads for different tasks. Thus, we have two distinct tasks/heads: (i) CoRoSeOf head: a 5-valued output trained with cross–entropy loss; (ii) REDv2 head: a 7-valued

Model	Transformer Model	F1	Acc
Baseline	romanian-bert-cased	0.8784	0.88
Feature Eng.	romanian-bert-cased	0.8778	0.8835
Multi-task	romanian-bert-cased	**0.8798**	**0.8844**
Baseline	xlm-roberta-large	0.7983	0.8348
Feature Eng.	xlm-roberta-large	0.8117	0.8489
Multi-task	xlm-roberta-large	**0.8399**	**0.8614**

Table 1: Training results on the presented model architectures.

Model	Transformer Model	F1	Acc
Multi-task 0.5	romanian-bert-cased	0.8789	0.8834
Multi-task 0.5	xlm-roberta-large	0.81	0.8465

Table 2: Training results on the multi-task model architecture, with 50% ratio between datasets.

output trained with binary cross–entropy loss. The input for both heads is, as in the case of the previous architectures, the pooled layer from the transformer model.

As a result of the different tweet counts contained in REDv2 and CoRoSeOf datasets, we had to batch them separately during the training process. Thus, at each step, we randomly pick one of the two datasets and create a batch with tweets from one of them, forwarding them up to the corresponding head. The datasets differ substantially in terms of size (CoRoSeOf contains 37.738 texts, whereas REDv2 contains 5.449), therefore we determine a probability threshold of 13.54% of the model choosing a batch from either dataset. In this manner, the model is able to see that the datasets are equally represented according to their size, and does not overfit on the smaller dataset. To verify this, we attempted to examine what the results would be if we forced the ratio to be 50/50. Our findings are illustrated in Table 2.

5. Experiments and Results

Our experiments were performed using the architectures described in Section 4, for all tested models, with a batch size of 16 (including for larger models). We tested more Romanian bert-base models, but for brevity only report the best performing one; we also test the xlm-roberta-large as a strong multilingual contender. Results are reported in Table 1, averaged across 5 runs.

To our expectations, the multi-task model achieves the highest F1-score and accuracy on both romanian-bert-cased and xlm-roberta-large. Between these two transformer models, romanian-bert outperforms the multilingual xlm-roberta by 4% F1-score and 2% accuracy, even if the roberta model is almost 3x larger (355M vs 124M), showing the power of monolingual models.

For variability, we have also trained the multi-task model using a ratio of 50% between both datasets. This means that at training time, the same amount of batches are taken from both CoRoSeOf and REDv2 datasets. It can be seen in Table 2 that using this ratio, both F1-score and accuracy, on both transformer models, are lower than the results in Table 1 when using the default ratio, which considers the disproportion between datasets when assigning training batches.

To compute a confidence interval of the results, we have trained the best performing model, the emotion and sexist and offensive language multi-task model which uses romanian-bert-cased, for a number of 20 times with random seeds. The averaged results are: F1-score 0.8791 and accuracy of 0.8827. The standard deviation for the F1-score is 0.0043, and for the accuracy is 0.0056. It can be observed that the standard deviation on both measures is small enough to consider these results reliable.

Class	Correct	Error	Support	Acc.
Non sexist	2916	141	3057	0.9539
Sexist direct	140	59	199	0.7035
Non sexist off.	210	173	383	0.5483
Sexist descriptive	64	56	120	0.5333
Sexist reporting	0	14	14	0

Table 3: Error analysis for each CoRoSeOf class.

5.1. Error Analysis

A detailed understanding of how failures occur or are distributed within the proposed models is essential. Thus, in Figure 3, we show the confusion matrix for one of our five experiments when creating the averaged multi-task model. This model has an overall accuracy of 0.8826. As can be seen in Figure 1, most of the test predictions belong to the neutral class, which is expected, since this is the class in CoRoSeOf that contains the most training texts.

Also, the *sexist reporting* class was never predicted in this experiment, which corroborates with the fact that this class contains only 140 texts and is also the smallest class in the CoRoSeOf dataset.

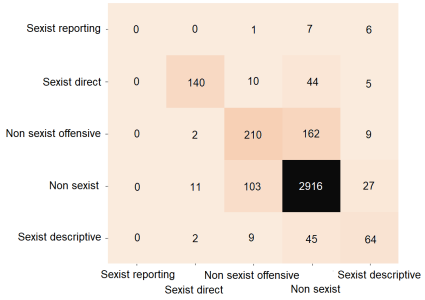

Figure 3: Confusion matrix for the romanian-bert-cased multi-task model

A detailed analysis of the errors is given in Table 3. The highest accuracy of 95.39% is achieved by the *non sexist* class, followed by *sexist direct* with 70.35%. The accuracy values of the *sexist descriptive* and *non sexist offensive* classes are rather similar, with 53.33% and 54.83%, respectively. In the case of *sexist reporting* class, the accuracy of the class is 0.

Analysing both Figure 3 and Table 3 the following phenomena stand out. The classification of *sexist reporting* tweets has never been accurate. The model considered seven of the texts to be *non sexist*, one to be *non sexist offensive*, and six to be *sexist descriptive*. In the example "era un moș beat mort în autobuz și se ținea în continuu după mine" ('there was a dead drunk old man on the bus and he kept following me'), the tweet is labeled as *sexist reporting*, however, the model classified it as *non sexist*. In addition, the emotion detection model indicates that this text reflects fear, which is an accurate prediction.

Sexist direct tweets were mainly classified correctly, with only 10 texts classified as *non sexist offensive*, 44 texts *non sexist*, and 5 texts *sexist descriptive*. The tweet, "bună, ce faci frumoaso" ('hello, how are you beautiful') was predicted as *non sexist* instead of *sexist direct*. In a surprising finding, the emotion detection model identified this text as joyful, showing that predicting the covert forms of sexism is challenging.

Two *non sexist offensive* tweets were miss-classified as *sexist direct*, 162 as *non sexist* and 9 as *sexist descriptive*. The text, "Judecătorii s-au șmecherit mult. Motivările cred că sunt lăsate in seama femeilor de serviciu..." ('The judges have become very sly. I think the motivations are left up to the maids...') was ini-

tially labelled as *non sexist offensive*, however, the model classified this text as *sexist descriptive*, which might be a true label in the opinion of some annotators.

The majority of *non sexist* tweets were classified correctly, although some exceptions were observed, including 11 tweets misclassified as *sexist direct*, 103 tweets misclassified as *non sexist offensive*, and 27 texts misclassified as *sexist descriptive*. The tweet, "Foarte frumoasă și sexy" ('Very beautiful and sexy') was classified as *sexist direct* by the model instead of the gold standard, *non sexist*. It is, however, apparent that the model generalizes correctly in this very subjective case. Furthermore, according to the emotion detection model, the emotion carried by this text is joy.

Finally, we noticed that the *sexist descriptive* tweets were mainly missclassified as *non sexist* (45 out of a total of 120), 2 *sexist direct*, and 9 *non sexist offensive*. The text "Trăiesc bărbat cu nevastă, darămite câine cu pisică" ('Man and wife live together, let alone dog and cat') is *sexist descriptive* but was misclassified as *non sexist*. The above is a classic example of how machine learning models are having difficulties to understand sarcasm, combined with the fact that our multi-task model has a preference for the *non sexist* class.

6. Conclusions and Future Work

We examine how emotions can be used to aid in categorizing sexist and offensive language; our experiments show that our proposed multi-task approach, addressed for the first time in the Romanian language for improving the detection of sexism and offensive language using emotions, is capable of producing convincing results. The dataset, code and results are freely available on GitHub.[4]

Further research will be carried out to test the effectiveness of the model in identifying sexist and offensive language using sarcasm and irony detection within texts as additional tasks. As a result of the imbalanced nature of CoRoSeOf, sampling data would be another approach to experiment with.

References

Abburi, Harika, Pulkit Parikh, Ni Chhaya, and Vasudeva Varma, 2020. Semi-supervised multi-task learning for multi-label fine-grained sexism classification.

Caruana, Rich, 1997. Multitask learning. *Machine Learning*, 28.

Ciobotaru, Alexandra, Mihai V. Constantinescu, Liviu P. Dinu, and Stefan Daniel Dumitrescu, 2022. RED v2: Enhancing RED Dataset for Multi-Label Emotion Detection. Marseille, France: European Language Resources Association (ELRA).

del Arco, Flor Miriam Plaza, M. Dolores Molina-González, Luis Alfonso Ureña López, and María Teresa Martín-Valdivia, 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

del Arco, Flor Miriam Plaza, María Dolores Molina-González, Luis Alfonso Ureña López, and María Teresa Martín Valdivia, 2022. Exploring the use of different linguistic phenomena for sexism identification in social networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022*, volume 3202 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[4] https://github.com/DianaHoefels/LTC23-MultiTaskLearning

Dumitrescu, Stefan, Andrei-Marius Avram, and Sampo Pyysalo, 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics.

Dumitrescu, Stefan Daniel and Andrei-Marius Avram, 2019. Introducing ronec–the romanian named entity corpus. *arXiv preprint arXiv:1909.01247*.

Hoefels, Diana Constantina, Çağrı Çöltekin, and Irina Diana Mădroane, 2022. CoRoSeOf - an annotated corpus of Romanian sexist and offensive tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association.

Jay, T., 1992. *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets*. Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards and on the Streets. J. Benjamins Publishing Company.

Lex, Alexander, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister, 2014. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko, 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics.

Pauwels, Anne, 2003. *Linguistic Sexism and Feminist Linguistic Activism*, chapter 24. John Wiley Sons, Ltd, pages 550–570.

Rajamanickam, Santhosh, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova, 2020. Joint modelling of emotion and abusive language detection.

Rodríguez-Sánchez, Francisco, Jorge Carrillo de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso, 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 69(0):229–240.

Ryosuke Nishi, Taro Takaguchi Keigo Oka Takanori Maehara Masashi Toyoda Ken-ichi Kawarabayashi Naoki Masuda, 2016. Reply trees in twitter: data analysis and branching process models. *Social Network Analysis and Mining*:1869–5469.

Sharifirad, Sima, Borna Jafarpour, and Stan Matwin, 2019. How is your mood when writing sexist tweets? detecting the emotion type and intensity of emotion using natural language processing techniques.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang, 2019. Mitigating gender bias in natural language processing: Literature review.

Verniers, Catherine and Jorge Vala, 2018. Justifying gender discrimination in the workplace: The mediating role of motherhood myths. *PLOS ONE*, 13:e0190657.

Zhang, Yu and Qiang Yang, 2022. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.