Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4142–4148
Marseille, 20-25 June 2022

Misogyny and Aggressiveness Tend to Come Together and Together We Address Them

Arianna Muti, Francesco Fernicola and Alberto Barrón-Cedeño
Department of Interpreting and Translation
Università di Bologna
Forlì, Italy
arianna.muti2@unibo.it francesco.fernicola@studio.unibo.it a.barron@unibo.it

Abstract

We target the complementary binary tasks of identifying whether a tweet is misogynous and, if that is the case, whether it is also aggressive. We compare two ways to address these problems: one multi-class model that discriminates between all the classes at once: not misogynous, non aggressive-misogynous and aggressive-misogynous; as well as a cascaded approach where the binary classification is carried out separately (misogynous vs non-misogynous and aggressive vs non-aggressive) and then joined together. For the latter, two training and three testing scenarios are considered. Our models are built on top of AlBERTo and are evaluated on the framework of Evalita's 2020 shared task on automatic misogyny and aggressiveness identification in Italian tweets. Our cascaded models —including the strong naïve baseline— outperform significantly the top submissions to Evalita, reaching state-of-the-art performance without relying on any external information.
Keywords: misogyny identification, aggressiveness identification, social media analysis

1. Introduction

Even if Twitter was conceived to express personal opinions, share big events, follow people or just communicate with friends, an increasing number of users misuse it by engaging in trolling, cyberbullying, or by posting aggressive and misogynous contents (Samghabadi et al., 2020). According to Nocentini et al. (2010), these contents feature an imbalance of power, intention, repetition, anonymity and publicity. Anonymity in particular has allowed the amount of hateful posts to dramatically increase. As a consequence, Twitter (as other platforms) struggles to control *inappropriate* contents.[1] According to Vox (Italian Rights Observer), women are more targeted than ever on Twitter. Although the overall number of hateful tweets decreased in 2020, the number of misogynous tweets increased significantly: from 26% to 49%.[2] Misogynous contents express hating towards women, in the form of insults, sexual harassment, male privilege, patriarchy, gender discrimination, belittling, violence, body shaming and sexual objectification (Srivastava et al., 2017). A misogynous text turns aggressive when it overtly or covertly encourages or legitimizes violence against women.

Since both Fulper et al. (2014) and Blake et al. (2021) showed that the amount of misogynous tweets is strongly correlated with rape and violence, employing automatic tools to identify them might limit those events. The development of tools to flag hateful language automatically —in particular towards women, since they are subjected to more bullying, abuse, hateful language, and threats than men— is a must (Fallows, 2005).

For these reasons, we approach the problem of spotting whether a tweet in Italian is misogynous and, if it is, whether it is also aggressive. We specifically focus on detecting aggressiveness in misogynous tweets. Existing systems struggle to identify the target of an aggressive statement, leading to a great number of false positives (Fersini et al., 2020). Therefore, with our experiments and our analysis, we hope to shed light on why existing models struggle to recognise the target of aggressive instances.

We explore two ways of addressing the two binary subtasks together: one cascaded combination of two binary models and one multi-class model that classifies instances into aggressive-misogynous vs misogynous vs other (Muti and Barrón-Cedeño, 2020). The two architectures, built on top of AlBERTo (Polignano et al., 2019), are compared over different scenarios against one naïve alternative which addresses the two sub-tasks independently and against the top Evalita 2020 AMI shared task models (Fersini et al., 2020).[3]

Our results show that addressing the two sub-tasks together is the best alternative —with the cascaded architecture performing better than the multi-class one— achieving state-of-the-art performance.

2. Related Work

Research on automatic misogyny identification on social media, in particular on Twitter, has been widely addressed with the AMI family of shared tasks, launched in the context of the EVALITA (Fersini et al., 2018a) and the IberEval (Fersini et al., 2018b) evaluation campaigns.

IberEval 2018 hosted a shared task on English and Spanish tweets (Fersini et al., 2018b).

[1] https://business.twitter.com/en/help/ads-policies/ads-content-policies/inappropriate-content.html

[2] http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-5/

[3] Evalita is a collection of shared tasks for the evaluation of NLP and speech tools for Italian (http://www.evalita.it); AMI stands for automatic misogyny identification.

Task A targeted misogyny identification and Task B presented a multi-class setting to classify sentences into seven misogyny categories: discredit, stereotype, objectification, threats of violence, dominance, derailing, and sexual harassment.

Participants mostly used representations based on n-grams and word embeddings, along with shallow models like support vector machines (SVM) and linear regression. Pamungkas et al., (2018) obtained the best score using an SVM including lexical features, although a purely lexical approach has proven to be suboptimal, as it fails to account for offensive words which may not necessarily target women (Ahluwalia et al., 2018).

The first edition of the AMI shared task for misogyny identification in Italian took place in 2018 (Fersini et al., 2018a). Task A addressed misogyny identification, while Task B aimed at recognising whether a misogynous tweet is person-specific or is generally addressed towards a group of women, additionally classifying the positive instances in the aforementioned seven categories. Bakarov (2018) obtained the best performance using TF-IDF weighting combined with singular value decomposition for representation and an ensemble of classifiers. The second edition took place in 2020 (Fersini et al., 2020) and focused on both misogyny and aggressiveness identification (cf. Section 3). Team unibo (Muti and Barrón-Cedeño, 2020) obtained the best performance with a multi-class approach, that we adopt as our `multi` approach (cf. Section 4). Team jigsaw (Lees et al., 2020) used an unconstrained setting, by augmenting the dataset with sentences sampled from the Italian Wikipedia articles that contain women-related identity terms, for a total of 11,000 sentences. They based their classifier on ensembles of fine-tuned custom BERT models, becoming the first runner-up. We compare our models against both unibo and jigsaw.

Another important contribution to misogyny detection was given by HatEval at SemEval 2019 (Basile et al., 2019), which focused on the detection of hate speech towards women and immigrants on Twitter in English and Spanish. Mozafar et al. (2019) showed that fine-tuning BERT achieves state-of-the-art results in this task. Another relevant shared task is sEXism Identification in Social neTworks (EXIST) (Rodríguez-Sánchez et al., 2021), which focuses on identifying sexism in tweets and gabs. Here misogyny is treated as a subcategory of sexism.

From a computational perspective, little distinction is made between the terms *sexism* and *misogyny* and usually researchers focus only on one of them. One exception is Parikh et al. (2021), who address both problems by first classifying sexism into 14 overlapping categories and then examining the efficacy of adapting the model for the detection and classification of misogynous tweets. According to them, sexism refers to discrimination on the basis of one's gender that predominantly afflicts women, whereas misogyny implies hate or entrenched prejudices against women. From a sociolinguistic point of view, sexism is described as the expression of male superiority over women, and misogyny as a more violent expression of sexism that implies hate (Savigny, 2020), sharing the view with Rodríguez-Sánchez (2021) who treats misogyny as a subcategory of sexism. Here we focus on misogyny only, according to the definition provided in Section 1.

Among other relevant works, Jaki et al. (2019) analysed and automatically identified misogynistic language on Incels.me, a forum created by male supremacists to disparage women (now shutdown). Fersini et al. (2020) showed that exploiting stilometry to profile users can lead to good discrimination of misogynous and not misogynous contents. Jha and Mamidi (2017) focused on identifying benevolent and hostile sexism, with the former being more subtle and the latter characterized by an explicitly negative attitude.

partition	misogynous (aggr.)	other	total
training	2,337 (1,783)	2,663	5,000
test	500 (176)	500	1,000

Table 1: Statistics of the AMI corpus for misogyny and aggressiveness identification.

3. Evaluation Framework

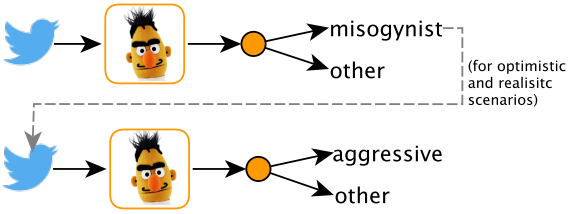
We perform our experiments on the evaluation framework of AMI 2020 Task A. AMI provides a corpus of $6,000$ tweets, manually annotated in two stages: (*i*) tweets are labelled as misogynous or not and (*ii*) involving only instances previously flagged as misogynous, tweets are labeled as aggressive or not. Table 1 shows some statistics. The training set is fairly balanced in terms of misogyny, and 76% of the misogynous tweets are aggressive. In the test set, 35% of misogynous instances are aggressive. We randomly split the training set into $4,700$ instances for training and 300 for tuning purposes. Models are evaluated by computing the average of the misogyny and aggressiveness tasks' F_1-measure.[4] We conduct a paired Student's T-Test (Dietterich, 1998) to assess if the models are statistically different, and run it against state-of-the-art approaches.

4. Models Description

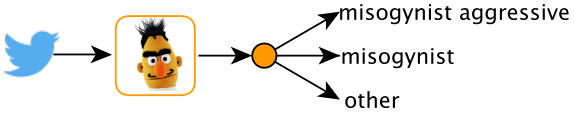
Our approach departs from Lees et al. (2020), since we adopt a constrained approach, neglecting external data. We build our models on top of AlBERTo (Polignano et al., 2019), a version of BERT (Devlin et al., 2019) trained on $200M$ Italian tweets including emojis, links, hashtags, and mentions. We use the pre-trained AlBERTo tokenizer to pre-process the text. Then we instantiate Pytorch's AlBERTo-Base, Italian Twitter lower cased[5], and fine-tune it on the downstream tasks. We use a softmax output layer with either two units for

[4]Evaluation code: `github.com/dnozza/ami2020`.

[5]`https://github.com/marcopoli/AlBERTo-it`.



(a) Cascaded architecture with two binary models (exps. `sing A` and `sing B`).



(b) Multi-class architecture (exp. `multi`).

Figure 1: The two system architectures.

the binary tasks of misogynous vs not-misogynous and aggressive vs not aggressive, or three units for the multi-class task aggressive-misogynous vs misogynous vs other. We compute independent losses for misogyny and aggressiveness in the cascaded `sing A` and `sing B` settings and one single loss for the multi-class `multi` settings, using the categorical cross-entropy loss.

We use the AdamW optimizer with $\epsilon = 1^{-8}$ (Loshchilov and Hutter, 2017) and greedily search for the optimal batch size and epoch number with a held-out strategy in ranges $[16, 32]$ and $[5, 8, 10, 15, 20]$ respectively. For testing, we re-train considering all the training material after finding the best parameters. All the experiments were run using Google Colab's GPU.

In addition to the previous models, we also experiment with multi-task learning in order to link both label representations together (Caruana, 1997). Refer to Appendix A for the model description and results.

5. Experiments and Results

Our objective is finding the best pipeline to discriminate misogynous and aggressive tweets from the rest. We draw three hypotheses. **H1:** Training the aggressiveness model on the full training set, even if half of the instances have not been judged and are assumed as non-aggressive (cf. Section 3), boosts the performance. The intuition is that, even if the extra instances are noisy, the more data the better. **H2:** Both tasks are interdependent and provide relevant information for each other, hence instantiating a cascaded model to solve one problem next to the other is better than addressing both independently. **H3:** A multi-class model performs better than a cascaded model because it assesses both problems at once, as Muti and Barrón-Cedeño (2020) claim.

We performed Experiments `sing A` and `sing B` to challenge H1. Both use the same architecture: two cascaded binary models, one for misogyny and one for aggressiveness, as shown in Figure 1(a). The difference lies in the training set for the aggressiveness model: in `sing A` both misogyny and aggressiveness models are trained on the whole dataset; in `sing B` the aggressiveness model is trained only on instances labeled as misogynous in the first place. Setting `sing B` aims at observing the behaviour of the aggressiveness model when neglecting potentially noisy non-misogynous instances. These settings intend to mimic the intuitive procedure when addressing the general problem: identifying if a tweet is misogynous and then, if true, whether it is also aggressive. This strategy mimics the corpus annotation process (cf. Section 3).

Table 2 shows the results; we focus on `sing A` vs `sing B` . As expected, the misogyny scores remain the same. Regardless of the scenario (used for H2 next), the aggressiveness model performs much better when `sing A` learns from the full dataset, despite the noise. The differences between these models are statistically significant (cf. Fig. 2). Therefore, we consider hypothesis H1 to be true: using the full (partially-noisy) training material allows for better generalization.

We consider three evaluation scenarios for `sing A` and `sing B` to test H2: (*i*) naïve (baseline): the aggressiveness model predicts on all instances, regardless of the misogyny information; (*ii*) optimistic: only instances labeled as misogynous according to the gold standard are assessed for aggressiveness (i.e. we assume a perfect misogyny classifier); and (*iii*) realistic: only instances identified as misogynous by our first classifier are classified by the aggressiveness model. For the last two scenarios all non-misogynous instances are flagged as non-aggressive.

Table 2 shows the results. Regardless of the quality of the misogyny classifier, the aggressiveness one benefits from the filtered input of the optimistic and realistic scenarios. It is worth noting that the performance shift is much smaller for the optimistic scenario (1.70 points absolute), in which 500 instances are simply assumed as non-aggressive. The drop in the naïve scenario is much bigger: 25.60 points absolute. Indeed, when comparing the three scenarios for both experiments, the naïve one —the only one which does not cascade the aggressiveness decision after that of misogyny—, consistently obtains the worst performance. This is a reflection of the veracity of H2: the better the input produced by the misogynous model (assumed to be perfect in the optimistic scenario), the more accurate the prediction of its aggressiveness. Once again, the differences are statistically significant according to the T-Test (cf. Fig. 2). Connecting the two problems together does matter, confirming our H2.

To challenge H3, model `multi` implements the top Evalita 2020 system of unibo (Muti and Barrón-Cedeño, 2020). So far, the evidence shows that a multi-class model discriminating aggressive–misogynous vs misogynous vs other at the same time generalizes better. Figure 1(b) represents its pipeline.

Table 2 shows the results; we now focus on `multi` . Considering the information on aggressiveness helps to improve the decisions on the misogyny class, lifting it

		misogyny		aggressiveness		overall
exp	**scenario**	**dev**	**test**	**dev**	**test**	**test**
singA	naïve	92.01	82.33	87.62	70.38	76.36
singA	optim.	92.01	82.33	87.62	75.66	79.00
singA	realistic	92.01	82.33	87.62	71.17	**76.75**
singB	naïve	92.01	82.33	75.25	44.78	63.56
singB	optim.	92.01	82.33	75.25	73.36	77.84
singB	realistic	92.01	82.33	75.25	64.94	73.64
multi	–	87.59	82.48	84.76	68.61	75.54
unibo	(Muti and Barrón-Cedeño, 2020)					74.38
jigsaw$_1$	(Lees et al., 2020)					74.06
jigsaw$_2$	(Lees et al., 2020)					73.80

Table 2: F_1 for singA , singB , multi and top Evalita 2020 models. We also include the three scenarios for aggressiveness detection for singA and singB . Average performance for three runs is displayed. The best configurations involve training during 8 epochs with a learning rate of 1-5, with a batch size of 4 for the single and 16 for the multi models.

by 0.15 points absolute. Nevertheless, this comes at the cost of a degradation in the prediction of aggressiveness, which drops by 3.86 points, causing the multi architecture to run short with an overall $F_1 = 75.54$ (1.21 lower than singA realistic). Thus, modelling the problem in a cascaded rather than a multi-class fashion is better, refuting H3. Nevertheless, both architectures could be combined: a cascaded model departing from the enhanced misogynous decision from multi and ending in singA for the aggressiveness decision results in $F_1 = 79.07$ in the optimistic and 76.83 in the realistic scenario. For comparison, the bottom of Table 2 shows the top Evalita 2020 models (Fersini et al., 2020). Our multi re-implementation of unibo (Muti and Barrón-Cedeño, 2020) outperforms it slightly, but the difference is not statistically significant (cf. Fig. 2). Our realistic singA model outperforms all Evalita systems, with statistically significant differences and reaching state-of-the-art performance even by neglecting any external data.

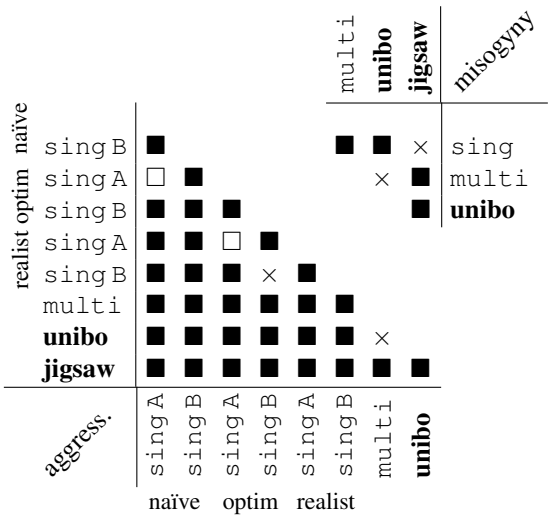


Figure 2: Student paired T-test for misogyny (top-right) and aggressiveness (bottom-left). Here, sing includes both singA and singB . Statistically-different models: ■→ with p-value< 0.01; □→with p-value< 0.05; × → not statistically different.

6. Error Analysis

We perform a manual analysis on the non-misogynous instances judged as aggressive in the naïve scenarios of singA and singB , which flag on average 11 and 258 non-misogynous instances as aggressive, respectively. While in the optimistic and realistic models they are assumed as non-aggressive, in the naïve model they are subject to the model decision.

Models based on singB rely on less negative instances, leading them to produce more false positives. In 80% of the cases, our predictions judged as false positives are indeed aggressive, but not according to the gold standard, because they are not women-oriented, while the final aim of this task is to spot aggressiveness targeting women. This is the case of instances 2 and 3 in Table 3. Words typically associated with aggression, such as *gola* (throat) and *schiaffi* (slaps) appear frequently in these instances (see instances 1 and 3 in Table 3). In both the training and test sets, *gola* is often the object of a violent act related to a sexual assault. All other instances contain swear words typical of aggressive language, in these cases used in non-aggressive circumstances and therefore misclassified as aggressive (see instance 4 in Table 3). Kurrek (2020) and Holgate (2018) show that, in hate speech identification, the presence of swear words often leads to false positives when they occur in non-abusive contexts. They are often misclassified even if their function is not harmful, as they serve to intensify emotions and sarcasm (Pamungkas et al., 2020). Some neutral instances are misclassified because they contain words that are prone to misinterpretation, as they normally occur in women-oriented aggression (*balena: whale/fat woman*, *scopo: aim/to f%ck*).

The word *scopo* occurs 42 times in the training set, solely in instances labeled as misogynous-aggressive,

	tweet	misogynous		aggressive	
		actual	pred	actual	pred
1	mamma tranquilla che non sono l'unica 20enne che non sa cosa cazzo deve fare della sua vita chill che è già bello che non mi sia **sparata in gola** anni fa [*chill mom I'm not the only 20-something who doesn't know what the fuck to do with her life just be thankful that I haven't* ***shot myself in the throat*** *years ago*]	no	no	no	yes
2	@Nigagalsen @matteosalvinimi un follower dal 2016…e ti permetti di aprire quella **lurida** fogna di bocca che hai. [*@Nigagalsen @matteosalvinimi following since 2016…and you dare open that* ***filthy*** *ass mouth of yours.*]	no	no	no	yes
3	Comunque sti uagliuncelli del cazzo che mettono la musica sotto la finestra della camera in cui dorme mio padre (tornato stanchissimo da lavoro come sempre) **li prenderei a schiaffi** uno a uno [***I really want to smack*** *each one of these fucking kids putting on music right under the window of the room where my father's sleeping (exhausted after his day's work, as usual)*]	no	no	no	yes
4	E bravo Ruggeri, finalmente ti sei ricordato che nel 77 eri il cantante dei #decibel, all'incirca mi ricordo qualche riga dei suoi testi ....."Che bello il lavaggio del cervello" Bum splash , la testa fa crash, **puttana** questo si chiama flash [*Look at you Ruggieri, you finally remembered that back in '77 you sang for the #decibel. I remember a few lyrics....."It's so nice to be brainwashed" Boom, splash! The head crashes!* ***Holy shit*** *now that's flash*]	no	no	no	yes

Table 3: Instances from the test partition, their actual class and the one predicted by our naïve models for both misogyny and aggressiveness. English translations included in italics.

it is therefore no surprise that our model classifies instances including that word as aggressive. However, this does not apply to *balena* which is equally used as an insult and as a reference to the animal.

We analyzed 300 random instances predicted by `multi` and focused again on aggressiveness, because this is where it struggles the most. The model tends to identify instances that contain verbs expressing an aggressive attitude, although not targeting women, as aggressive. We find that most of the instances refer to violent acts related to the throat and slaps. Other instances classified as such are neither misogynous nor aggressive, but do contain semantically ambiguous words used to insult women (e.g., *acida*: acid, peevish).

7. Conclusions and Further Work

We presented a number of architectures for the tasks of misogyny and aggressiveness identification in Italian tweets based on AlBERTo, and evaluated them on the 2020 edition of the AMI shared task. Our experiments showed that addressing the two problems together —through two cascaded binary models— results in the best performance, even without relying on any additional information. The top model obtains $F_1 = 76.75$ and significantly outperforms our baselines and the best AMI 2020 shared task systems (even those which employ external information, with maximum $F_1 = 74.38$). However, our models struggle the most when identifying aggressiveness, confirming that it is hard to identify the target gender of an aggressive statement.

In the future, we plan to adopt an unconstrained approach and increase the number of training instances. It would be interesting to extend the scope of the model as well and annotate the training set according to the gender of the target (for instance, by means of mentions). This might help discriminate better between women- and men-oriented aggressive language, which is a timely problem and was one of the biggest sources of difficulty for our models. We also plan to extend this experiment to other languages, such as English and Spanish.

8. Acknowledgements

We thank Lees et al. (2020) for sharing their runs to perform t-tests against their models.

A. Muti's research is carried out under project "DL4AMI–Deep Learning models for Automatic Misogyny Identification", in the framework of *Progetti di formazione per la ricerca: Big Data per una regione europea più ecologica, digitale e resiliente*—Alma Mater Studiorum–Università di Bologna, Ref. 2021-15854.

9. Bibliographical References

Ahluwalia, R., Shcherbinina, E., Callow, E., Nascimento, A., and De Cock, M. (2018). Detecting misogynous tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-*

located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), page 2150:242–248, Sevilla, Spain. CEUR-WS.org.

Bakarov, A. (2018). Vector space models for automatic misogyny identification. In *Proceedings of the sixth evaluation campaign of natural language processing and speech tools for Italian. Final workshop (EVALITA 2018) co-located with the fifth Italian conference on computational linguistics (clic-it 2018)*. CEUR-WS.org.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Basile, V., Croce, D., Maro, M. D., and Passaro, L. C. (2020). EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, volume 2765. CEUR-WS.org.

Blake, K., O'Dean, S., Lian, J., and Denson, T. (2021). Misogynistic tweets correlate with violence against women. *Psychological Science*, 32:095679762096852, 02.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Fallows, D. (2005). How women and men use the internet. Technical report, Pew Internet & American Life Project, December.

Fersini, E., Nozza, D., and Boifava, G. (2020). Profiling Italian misogynist: An empirical study. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France, May. European Language Resources Association (ELRA).

Fulper, R., Ciampaglia, G. L., Ferrara, E., Ahn, Y.-Y., Flammini, A., Menczer, F., Rowe, K., and Lewis, B. (2014). Misogynistic language on twitter and sexual violence. In *Proceedings of the acm web science workshop on computational approaches to social modeling (chasm)1–4.*

Holgate, E., Cachola, I., Preoţiuc-Pietro, D., and Li, J. J. (2018). Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4405–4414, Brussels, Belgium, October-November. Association for Computational Linguistics.*

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., and De Pauw, G. (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2):240–268. Publisher: John Benjamins Type: Journal Article.

Jha, A. and Mamidi, R. (2017). 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*, pages 7–16, Vancouver, Canada, August. Association for Computational Linguistics.

Kurrek, J., Saleem, H. M., and Ruths, D. (2020). Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. *Proceedings of the Fourth Workshop on Online Abuse and Harms (pp. 138-149)*, arXiv:1503.06733.

Lees, A., Sorensen, J., and Kivlichan, I. (2020). Jigsaw@AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)* (Basile et al., 2020).

Loshchilov, I. and Hutter, F. (2017). Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection in online social media. In *Proceedings of the International Conference on Complex Networks and Their Applications*. Springer.

Muti, A. and Barrón-Cedeño, A. (2020). UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)* (Basile et al., 2020).

Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., and Menesini, E. (2010). Cyberbullying: Labels, behaviours and definition in three european countries. *Australian Journal of Guidance and Counselling*, 20(2):129–142.

Pamungkas, E. W., Cignarella, A. T., Basile, V., and Patti, V. (2018). 14-ExLab@UniTo for AMI at IberEeal2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *IberEval@SEPLN*.

Pamungkas, E. W., Basile, V., and Patti, V. (2020). Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57.

Parikh, P., Abburi, H., Chhaya, N., Gupta, M., and Varma, V. (2021). Categorizing sexism and misogyny through neural approaches. *ACM Trans. Web*, 15(4), jun.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., and Donoso, T. (2021). Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

Samghabadi, N. S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., and Solorio, T. (2020). Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Savigny, H. (2020). *Sexism and Misogyny*. John Wiley & Sons, Ltd.

Srivastava, K., Chaudhury, S., Bhat, P., and Sahu, S. (2017). Misogyny, feminism, and sexual harassment. *Industrial psychiatry journal*, 26(2):111–113.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

10. Language Resource References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, June. ACL.

Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the Evalita 2018 task on automatic misogyny identification (AMI). In *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples*, pages 59–66. Torino: Accademia University Press.

Fersini, E., Rosso, P., and Anzovino, M. (2018b). Overview of the task on automatic misogyny identification at IberEval 2018. In *Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, Sevilla, Spain.

Fersini, Elisabetta and Nozza, Debora and Rosso, Paolo. (2020). *AMI@EVALITA2020: Automatic Misogyny Identification*. CEUR.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481, Bari, Italy. CEUR.

A. Multi-Task Approach

For the multi-task approach we adopt a training method where a single shared encoder is mapped across the different tasks, meaning that two separate replicas are created, sharing the same internal parameters. Once the backpropagation step is performed, the weights of both encoders are updated in parallel, transferring the information learned in one task to the other. For this scenario we build our model on top of XLM-RoBERTa-Base (Wolf et al., 2020).[6] We adopt a learning rate of 2e-5 and use the AdamW optimizer (Hu et al., 2020). We explore the epochs $\in [1, 4, 8]$ and test an effective training batch size of 2. In this case the training was carried out using a NVIDIA Quadro P4000 8GB GPU. The performance is significantly lower than all other models, achieving an overall $F_1 = 64.83$ at 4 epochs, before degrading to an $F_1 = 62.42$ at 8 epochs, falling short of more than 10 points with respect to `singA`. While the prediction of misoginy achieves an $F_1 = 71.34$, the biggest loss in performance lies once again in the prediction of aggressiveness, which achieves an $F_1 = 58.33$ which, while not being as low as `singB` naïve, it still constitutes the second lowest performance, with almost 15 points less than `singA`.

This further confirms the results of the other experiments, where the prediction of aggressiveness was once again the most challenging aspect. Additionally, this provides insight on the importance of directionality of the architecture. This is due to the fact that in the multi-task model both encoders are trained at the same time and it does not cascade the aggressiveness decision after that of misoginy, similarly to the naïve scenario. Nevertheless, this model does achieve a higher performance in terms of aggressiveness prediction, thus further experiments are advisable in the multi-task setting to explore this aspect by implementing a cascaded effect from one task to the other to further assess the impact of directionality on the predictions.

[6] `https://huggingface.co/xlm-roberta-base`