Hate Speech Detection in an Italian Incel Forum Using Bilingual Data for Pre-Training and Fine-Tuning

Paolo Gajo[1], Silvia Bernardini[1], Adriano Ferraresi[1] and Alberto Barrón-Cedeño[1]

[1]Department of Interpreting and Translation, Università di Bologna, Corso della Repubblica, 136, 47121, Forlì, FC, Italy

Abstract

English. In this study, we aim to enhance hate speech detection in Italian incel posts. We pre-train monolingual (Italian) and multilingual Transformer models on corpora built from two incel forums, one in Italian and one in English, using masked language modeling. Then, we fine-tune the models on combinations of English and Italian corpora, annotated for hate speech. Experiments on a hate speech corpus derived from the Italian incel forum show that the best results are achieved by training multilingual models on bilingual data, rather than training monolingual models on Italian-only data. This emphasizes the importance of using training and testing data from a similar linguistic domain, even when the languages differ.

Italiano. *In questo studio, ci proponiamo di migliorare il rilevamento dei discorsi d'odio in post tratti da un forum italiano di incel. Addestriamo modelli Transformer mono (italiano) e multilingue su corpora ottenuti da due forum di incel, uno in italiano e uno in inglese, con il masked language modeling. Facciamo quindi il fine-tuning dei modelli su corpora in italiano e inglese con annotazioni indicanti se un post esprime odio. Sperimentando su un corpus annotato per i discorsi di odio ottenuto da un forum italiano di incel mostriamo che i risultati migliori si ottengono addestrando modelli multilingue su combinazioni bilingue di corpora e non con modelli italiani e dati monolingue. Ciò sottolinea l'importanza di utilizzare dati di addestramento appartenenti a un contesto linguistico simile a quello dei dati di valutazione, anche con lingue differenti.*

Keywords

incels, hate speech, masked language modeling, transformers, bert, multilingual mlm, multilingual masked language modeling

1. Introduction

While there is no scarcity of English-language models and training resources for the detection of hate speech (HS), especially with the recent rise in popularity of this research topic [1], much work can still be carried out on this problem in other languages. For less-resourced languages, such as Italian, one of the main difficulties of combating this phenomenon is the lack of annotated data [2]. The problem is even more severe when considering the detection of hate speech in niche contexts, such as in forums frequented by incels, short for "involuntary celibates", a community known for its hateful language [3, 4] and use of specific misogynous and racist lexicon [5, 6]. In particular, it seems no work has yet been done on the detection of hate speech in Italian incel forums.

In this paper, we present a simple approach to improve the performance of hate speech detection models in Italian forums frequented by incels. Our contribution is two-fold:
***(i)* Masked language modeling.** We adapt monolingual Italian models to the linguistic domain of Italian incel forums by training them on the masked language modeling (MLM) task. As training material, we use an unlabelled corpus compiled from an Italian incel forum. We also adopt an existing multilingual model, already domain-adapted to the incel domain in both English and Italian. We release these novel models, which can be used for further research on the topic.[1]
***(ii)* Hate speech detection.** We fine-tune the vanilla and domain-adapted models on the downstream task of detecting hate speech in Italian incel posts. Monolingual models are trained on Italian-only combinations of corpora binary-annotated for hate speech, while Italian–English combinations are used for the multilingual models.

Testing the performance of the models on a labelled hate speech corpus, obtained by annotating posts from the Italian incel forum, shows that the best results are obtained by first training the base multilingual model on bilingual data taken from both the Italian and English incel forums, using the MLM task, and then fine-tuning it on combinations of Italian and English corpora, annotated for hate speech. In the approached scenarios, pre-training and fine-tuning on bilingual in-domain incel annotated data may therefore be more effective than training on general target-language labelled corpora, despite part of the training data not being in the language of the downstream

[1]Links to the models, released on HuggingFace: https://github.com/paolo-gajo/clic23.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

paolo.gajo2@unibo.it (P. Gajo); silvia.bernardini@unibo.it (S. Bernardini); adriano.ferraresi@unibo.it (A. Ferraresi); a.barron@unibo.it (A. Barrón-Cedeño)
https://www.unibo.it/sitoweb/paolo.gajo2 (P. Gajo); https://www.unibo.it/sitoweb/silvia.bernardini (S. Bernardini); https://www.unibo.it/sitoweb/adriano.ferraresi (A. Ferraresi); https://www.unibo.it/sitoweb/a.barron (A. Barrón-Cedeño)
0009-0009-9372-3323 (P. Gajo); 0000-0003-0750-4861 (S. Bernardini); 0000-0002-6957-0605 (A. Ferraresi); 0000-0003-4719-3420 (A. Barrón-Cedeño)

© 2023 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)

task. In addition, the results show that this strategy can be used to improve model performance when in-domain target-language data is scarce, by using in-domain data from other languages.

The rest of the paper is organized as follows: Section 2 presents related work on hate speech detection in Italian and English, as well as multilingual approaches to the problem. Section 3 describes the corpora used in this study. Section 4 presents the employed models. Section 5 describes the experiments conducted and discusses the results. Section 6 closes the contribution with conclusions and future work.

2. Related Work

Prior work on Italian hate speech detection has been conducted chiefly within the context of EVALITA. The 2018 edition hosted a shared task on hate speech detection [7] based on two corpora, one comprising tweets and one Facebook posts. The participating teams experimented with a variety of algorithms, with the top team relying on an SVM and a BiLSTM [8]. The 2020 edition hosted a shared task on the detection of hate speech, stereotypes, and nominal utterances, especially against migrants, focusing on tweets and news headlines [9]. In this case, the best team's approach for the hate speech detection sub-task [10] was to fine-tune BERT$_{base}$ [11] along with AlBERTo [12] and UmBERTo [13], two BERT models pre-trained on Italian data.

As regards misogyny in particular, EVALITA 2018 hosted for the first time a shared task on automatic misogyny identification (AMI), where the top performing teams used a combination of TF-IDF and SVD for the Italian scenario, and TF-IDF with logistic regression for the English one [14]. EVALITA 2020 hosted the second edition of the AMI shared task, focusing on Italian tweets [15], where an ensemble of BERT models obtained the top performance [16]. In EVALITA 2023, Di Bonaventura et al. [17] used triple verbalisation, prompting and majority vote to improve the performance of an AlBERTo model on the tasks of homotransphobia and hate speech detection.

English-language hate speech detection has been conducted in a variety of ways. Among others, Davidson et al. [18] build a corpus of tweets annotated with multi-class labels ("hate speech", "offensive", "neither") and train logistic regression and linear SVM models on it. Mathew et al. [19] build a corpus called HateXplain from Twitter and Gab posts, annotated with multi-class labels based on whether the post is "offensive", expresses "hate", or is "normal", which they use to fine-tune a BERT hate speech classifier. Caselli et al. [20] retrain BERT$_{base}$ on the MLM task using an unlabelled corpus built from hateful and offensive Reddit messages, obtaining a model called HateBERT, capable of outperforming BERT$_{base}$ on hate speech identification on various benchmark datasets.

In multilingual settings, Pelicon et al. [21] use a multilingual combination of corpora annotated for hate speech to improve the performance of classifiers in zero-shot, few-shot and well-resourced settings. Gokhale et al. [22] use MLM training to improve the hate speech detection performance of BERT in Hindi and Marathi, separately. We follow such approaches in improving the performance of our models, with a specific focus on monolingual vs. bilingual pre-training, compared to Gajo et al. [23].

3. Corpora

We leverage three labelled Italian-language corpora from past EVALITA campaigns, along with two labelled corpora compiled from two incel forums.

EVALITA corpora The first Italian corpus we use was compiled for the first edition of the Hate Speech Detection (HaSpeeDe) shared task, from EVALITA 2018 [7] (henceforth HSD-FB), by annotating Facebook posts for hate speech. The second one is from the 2020 edition of HaSpeeDe [24] (HSD-TW), compiled by adding new data to the HaSpeeDe 2018 Twitter corpus. The third corpus is the one compiled for the Automatic Misogyny Identification (AMI) shared task [15] (AMI-20), hosted at EVALITA 2020. AMI-20 is annotated with misogyny labels, which we use as hate speech labels to train our classifiers. Where the corpora were not partitioned, we split them 70/30 between training and development sets. We do not use the test partitions, as we are interested in maintaining consistency with the use of the original splits of these corpora.

Incel corpora We use two unlabelled corpora compiled by scraping two incel forums [23]: *Incels.is*[2] and *Il forum dei brutti*[3], respectively in English and Italian.

A subset of the two corpora was annotated for both misogyny and racism.[4] The annotated partitions are referred to as IFS-EN and IFS-IT ("Incel Forum, Supervised, English" and "Italian").

We keep the training, development and testing partitions as in the released corpora. IFS-IT is used in its entirety solely as a test set, due to the unavailability of additional annotated Italian incel data for training. Said scarcity prompted us to leverage the available data in order to conduct cross-lingual experiments for the incel domain, for which Italian is a low-resource language.

[2] https://incels.is (Last access: 11 Aug 2023)
[3] https://ilforumdeibrutti.forumfree.it (Last access: 11 Aug 2023)
[4] Refer to Gajo et al. [23] for details on the annotation process.

Table 1
Existing corpora class distribution.

Corpus	HS	Non-HS
AMI-20 [15]	2,337	2,663
HSD-FB [7]	1,382	1,617
HSD-TW [7]	971	2,028

Table 1 shows the class distribution of all three EVALITA corpora, whereas Table 2 shows the distribution for the incel corpora, where posts are considered hateful if they are either labeled as misogynous or racist. As can be inferred from the statistics, while misogynous instances comprise around 39% of the instances in both IFS-EN and IFS-IT, the same cannot be said for the racist ones, which are much more prevalent in IFS-EN (13% vs. 0.03%). This shows a clear difference in terms of the hate speech produced by the two incel communities.

4. Models

With relation to the Italian-only scenario, we use UmBERTo and AlBERTo for our baseline models. We choose these models because they achieved the best performance in previous EVALITA shared tasks on hate speech [9] and misogyny [15] identification. In order to improve the performance of the two models on the task of identifying hate speech in Italian incel forums, we train them on the MLM task on posts extracted from *Il forum dei brutti*. We follow this approach because it has been shown to work in English both for general hateful content [20] and incel forums [23]. For training data, we use the entirety of the contents of the forum, for a total of $627k$ posts. The intersection between the unlabelled incel corpora and the annotated corpora listed in Table 2 is void. That is, none of the data contained in IFS-IT was obtained from the Italian data scraped from *Il forum dei brutti* and used for MLM pre-training. The same is true for IFS-EN and the English MLM pre-training data taken from *Incels.is*. Doing this, we obtain two new models which we refer to as "Incel UmBERTo" and "Incel AlBERTo".

The MLM pre-training process is carried out in all cases by tokenizing post contents using each model's own tokenizer and masking tokens with a probability of 15%. We use a batch size of 32 samples and train the models for one epoch on one Tesla P100 GPU with 16 GB of VRAM.

As regards the bilingual setting, we use mBERT$_{base}$ as our baseline. We also use an MLM-enhanced version of it, "Incel mBERT",[5] obtained by further pre-training mBERT$_{base}$ on $500k$ posts sampled from *Il forum dei brutti* and $500k$ posts sampled from *Incels.is*, for a total of $1M$ posts in Italian and English [23].

[5] https://huggingface.co/pgajo/incel-mbert

Table 2
Incel corpora class distribution [23].

Corpus	Misogyny	Racism	Both	Neither
IFS-EN$_{tr}$	806	630	46	2,160
IFS-EN$_{de}$	173	130	13	464
IFS-EN$_{te}$	160	125	7	489
IFS-IT$_{te}$	187	8	5	300

5. Experiments and Results

We approach the task of identifying hate speech as a binary classification problem, where a post can either be hateful or not. We train each model five times on all possible combinations of the corpora listed in Tables 1 and 2 in order to make our results more reliable and diminish the effect of the random initialization of the models. In the monolingual Italian setting we never use IFS-EN, while it is always included when training the multilingual models in the bilingual setting. We select the number of epochs based on the convergence of the performance on the validation set, in terms of F_1-measure on the positive class. For each corpus combination, the training and validation sets are the union of the individual training and validation sets of each merged corpus. The models are then evaluated on the IFS-IT test set.

Monolingual setting Table 3 shows the performance in terms of precision, recall and F_1-measure for the Italian-only models and corpora combinations. The top-performing model is Incel AlBERTo, which achieves a test F_1 of 0.707 when training solely on HSD-FB. Compared to AlBERTo, this represents an improvement of 2.4 points. To a lesser degree, the same can be observed with regard to Incel UmBERTo and UmBERTo (+0.9 F_1 points), when using the same combination. In both cases, this shows that pre-training AlBERTo and UmBERTo using MLM on Italian posts extracted from *Il forum dei brutti* is effective in improving their performance.

The worst results are obtained when training solely on HSD-TW, with Incel AlBERTo and Incel UmBERTo performing worse than UmBERTo and AlBERTo, showing an opposite trend to the one observed when training on HSD-FB. The validation scores are also lower for HSD-TW combinations, compared to combinations including HSD-FB, showing that the models have a harder time learning from HSD-TW. This is coherent with the results obtained by teams participating in the two HaSpeeDe shared tasks [7, 9] and with the fact that HSD-FB's messages are "longer and more correct than those in Twitter, allowing systems (and humans too) to find more and more clear indications of the presence of HS" [7]. The fact that messages in HSD-FB are longer is also coherent with the Italian incel models performing better than the

Table 3
Performance when fine-tuning the monolingual models on Italian-only corpora. Epochs (e) selected based on validation F_1. Best scores in bold, second-best underlined; ■ = corpus used for training.

	HSD-FB	HSD-TW	AMI-20	(e)	$\mathbf{F}_{1\ val}$	$\mathbf{R}_{val}$	$\mathbf{P}_{val}$	$\mathbf{F}_{1\ test}$	$\mathbf{R}_{test}$	$\mathbf{P}_{test}$
UmBERTo	■			5	0.855±0.003	0.868	0.843	0.696±0.010	**0.879**	0.576
		■		4	0.754±0.004	0.800	0.713	0.432±0.060	0.319	0.685
			■	4	0.914±0.004	0.931	**0.899**	0.569±0.031	0.520	0.631
	■	■		4	0.788±0.006	0.824	0.755	0.666±0.024	0.758	0.595
	■		■	5	0.883±0.004	0.900	0.867	0.697±0.019	0.747	0.653
		■	■	5	0.828±0.003	0.844	0.814	0.596±0.017	0.526	0.688
	■	■	■	5	0.822±0.003	0.836	0.808	0.680±0.016	0.692	0.671
Incel UmBERTo	■			5	0.867±0.006	0.887	0.848	**0.705±0.009**	0.870	0.593
		■		4	0.756±0.002	0.810	0.708	0.403±0.024	0.285	0.692
			■	4	**0.918±0.001**	**0.946**	0.891	0.652±0.031	0.608	0.705
	■	■		4	0.790±0.003	0.831	0.754	0.660±0.014	0.696	0.627
	■		■	5	0.886±0.002	0.901	0.872	0.704±0.005	0.732	0.678
		■	■	2	0.831±0.003	0.866	0.799	0.648±0.011	0.544	**0.802**
	■	■	■	5	0.828±0.003	0.853	0.804	0.699±0.029	0.718	0.682
AlBERTo	■			4	0.850±0.003	0.899	0.807	0.683±0.006	**0.941**	0.537
		■		1	0.752±0.006	0.817	0.698	0.520±0.089	0.426	0.716
			■	2	0.907±0.004	**0.952**	0.866	0.528±0.022	0.517	0.542
	■	■		2	0.775±0.003	0.803	0.750	0.695±0.007	0.786	0.623
	■		■	3	0.879±0.003	0.918	0.843	0.705±0.011	0.803	0.629
		■	■	3	0.820±0.001	0.888	0.762	0.652±0.018	0.645	0.660
	■	■	■	2	0.808±0.011	0.872	0.753	0.684±0.015	0.821	0.587
Incel AlBERTo	■			5	0.847±0.005	0.863	0.831	**0.707±0.007**	0.791	0.639
		■		1	0.748±0.002	0.785	0.715	0.506±0.035	0.370	**0.805**
			■	5	**0.912±0.003**	0.930	**0.895**	0.617±0.018	0.562	0.685
	■	■		2	0.771±0.004	0.791	0.752	0.673±0.016	0.721	0.632
	■		■	5	0.873±0.003	0.888	0.858	0.668±0.014	0.663	0.674
		■	■	1	0.818±0.004	0.864	0.776	0.656±0.007	0.593	0.736
	■	■	■	4	0.800±0.009	0.828	0.773	0.688±0.017	0.747	0.639

vanilla models when training on HSD-FB, since *Il forum dei brutti* on average contains rather long posts (∼53 avg. tokens),[6] unlike Twitter corpora, which were limited to 280 characters per tweet prior to 2023. Finally, another element which might explain the lower performance when training on HSD-TW is that it contains hate speech against migrants, which might not be as relevant when it comes to *Il forum dei brutti*, since racism is not all that prevalent in this forum, compared to misogyny.

As regards combining different Italian corpora, the strategy yields the highest performance for AlBERTo and UmBERTo when training on both HSD-FB and AMI-20. However, once the models are MLM-trained on *Il forum dei brutti*, the performance decreases for some combinations, with MLM pre-training seemingly nullifying the improvements obtained by merging different corpora. Therefore, while some improvement can be observed by merging different corpora, MLM appears to be a more effective strategy for improving the performance of the models, although it requires greater computational resources.

Bilingual setting Table 4 reports the results for the bilingual setting. Compared to the best combination using mBERT$_{base}$, which achieves a test F_1 of 0.688, the best combination using Incel mBERT achieves a test F_1 of 0.722 (+3.4 F_1 points), which is also the highest score across both language settings. Just like in the monolingual setting, mBERT$_{base}$ performs better when only training it on HSD-FB (in addition to IFS-EN). Conversely, Incel mBERT performs better when training on AMI-20 and IFS-EN. This is interesting, since the incorporation of the AMI-20 corpus lowered the performance of all Italian-only models, compared to only training on HSD-FB. Since misogyny is the main way hate speech is expressed in *Incels.is* (39.44% of the instances in IFS-EN are misogynous) and Incel mBERT was pre-trained using posts extracted from this forum, the performance boost could be due to the fact that the model is better at learning about misogynous language compared to mBERT$_{base}$ and the Italian-only models.

[6] Obtained with BertTokenizer: https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertTokenizer

Table 4
Performance when fine-tuning the multilingual models on mono- and bilingual corpora combinations. Epochs (e) selected based on validation F_1. Best scores in bold; ■ = corpus used for training.

		HSD-FB	HSD-TW	AMI-20	IFS-EN	(e)	$\mathbf{F}_{1\ val}$	$\mathbf{R}_{val}$	$\mathbf{P}_{val}$	$\mathbf{F}_{1\ test}$	$\mathbf{R}_{test}$	$\mathbf{P}_{test}$
mBERT	Monolingual	■				4	0.807±0.008	0.846	0.775	0.651±0.013	0.891	0.516
			■			4	0.717±0.012	0.745	0.693	0.493±0.037	0.455	0.540
				■		5	**0.885±0.003**	**0.890**	**0.881**	0.425±0.034	0.384	0.479
		■	■			2	0.732±0.014	0.743	0.724	0.619±0.034	0.711	0.552
		■		■		3	0.844±0.002	0.873	0.818	0.639±0.023	0.747	0.560
			■	■		3	0.789±0.008	0.830	0.754	0.560±0.023	0.621	0.516
		■	■	■		5	0.784±0.002	0.793	0.775	0.600±0.014	0.634	0.569
					■	5	0.846±0.010	0.854	0.837	0.465±0.046	0.345	**0.725**
	Bilingual	■			■	3	0.841±0.004	0.852	0.832	**0.688±0.006**	**0.921**	0.549
			■		■	3	0.846±0.002	0.866	0.827	0.529±0.041	0.482	0.588
				■	■	5	0.844±0.007	0.849	0.838	0.634±0.023	0.587	0.692
		■	■		■	4	0.844±0.006	0.844	0.844	0.616±0.028	0.675	0.568
		■		■	■	5	0.842±0.004	0.844	0.840	0.676±0.012	0.801	0.585
			■	■	■	4	0.847±0.008	0.841	0.853	0.570±0.048	0.535	0.620
		■	■	■	■	5	0.837±0.008	0.829	0.845	0.613±0.019	0.639	0.590
Incel mBERT	Monolingual	■				1	0.817±0.009	0.865	0.777	0.668±0.016	0.841	0.557
			■			5	0.727±0.006	0.757	0.701	0.461±0.032	0.379	0.589
				■		4	**0.895±0.007**	**0.912**	**0.879**	0.574±0.047	0.511	0.666
		■	■			3	0.747±0.007	0.774	0.723	0.605±0.015	0.640	0.574
		■		■		3	0.854±0.004	0.896	0.816	0.654±0.025	0.752	0.583
			■	■		4	0.802±0.003	0.840	0.767	0.645±0.019	0.655	0.639
		■	■	■		3	0.799±0.004	0.825	0.774	0.648±0.022	0.644	0.653
					■	3	0.855±0.003	0.877	0.834	0.516±0.071	0.386	**0.807**
	Bilingual	■			■	5	0.859±0.010	0.853	0.864	0.708±0.007	**0.889**	0.588
			■		■	2	0.861±0.015	0.866	0.856	0.615±0.025	0.558	0.690
				■	■	4	0.853±0.009	0.863	0.844	**0.722±0.028**	0.704	0.746
		■	■		■	5	0.857±0.007	0.855	0.859	0.679±0.014	0.731	0.635
		■		■	■	4	0.856±0.009	0.857	0.856	0.689±0.011	0.707	0.673
			■	■	■	5	0.850±0.007	0.839	0.860	0.644±0.010	0.580	0.725
		■	■	■	■	5	0.869±0.003	0.878	0.861	0.700±0.013	0.702	0.698

On average, the lowest performance is achieved when training separately on IFS-EN and on the Italian corpora (monolingual rows in Table 4). When using bilingual data, the worst results are obtained when training on HSD-TW and combinations containing it, coherently with the results in the monolingual settings shown in Table 3.

For almost all combinations of Italian corpora, performance increases once IFS-EN is added to the training data, i.e. bilingual data leads to better performance.

Monolingual vs. bilingual The results of our experiments show that the highest performance is not obtained by fine-tuning on the Italian-only corpus combinations, but on the bilingual ones. Indeed, for four bilingual corpus combinations out of seven, Incel mBERT's performance is higher than all other models. The combinations for which Incel mBERT does not beat all the others are HSD-FB+HSD-TW, HSD-FB+AMI-20 and HSD-TW+AMI-20.

Since mBERT was originally pre-trained in 104 languages and AlBERTo and UmBERTo were pre-trained only on Italian corpora, the fact that Incel mBERT can outperform them by pre-training on just $1M$ bilingual instances is rather unexpected. Even more interesting is the fact that, although we are testing on an entirely Italian corpus, Incel mBERT also outperforms Incel AlBERTo and Incel UmBERTo. Therefore, in the approached scenarios, using bilingual instances to pre-train a multilingual model using MLM yields higher performance than pre-training Italian models only on Italian posts. Furthermore, the number of Italian posts used to train Incel AlBERTo and Incel UmBERTo is $627k$, which is greater than the $500k$ Italian posts used for Incel mBERT.

As such, we could arguably conclude that the model is learning to spot hate speech more effectively in IFS-IT by learning language-agnostic incel concepts, since Incel mBERT is pre-trained on posts extracted from two incel forums in two different languages. Although the

Table 5
Performance of our best model on on the different subclasses of the IFS-IT test corpus.

Class	Precision	Recall	F_1	Inst.
None	0.806	0.857	0.830	300
Misogyny	0.746	0.674	0.708	187
Racism	1.000	0.875	0.933	8
Both	1.000	1.000	1.000	5
Macro	0.888	0.851	0.868	500

two considered incel communities are distinct, the hateful Red Pill ideology has spread internationally and is shared by both. This could explain why Incel mBERT performs better than the Italian-only models: the model might be learning about incel hate speech by paying more attention to the sociological concepts underlying the language, and putting less focus on purely linguistic features, ultimately improving its performance.

Performance on misogyny/racism subclasses
Table 5 reports the performance of the best model —the one obtained by fine-tuning Incel mBERT on IFS-EN ∪ AMI-20— on the individual misogyny and racism labels of the IFS-IT test set. When looking at the difference between the performance on misogyny vs. racism, we notice a stark difference, with racism having perfect precision and a much higher F_1. Expectedly, this also translates into the instances that are both misogynous and racist, with perfect precision and recall. The explanation for the racist instances being much easier to detect is two-fold: (*i*) the number of instances which are only racist is much smaller (8 vs. 187) and (*ii*) compared to the misogyny expressed by *Il forum dei brutti* users, the racism is much more explicit and simpler to identify. This can be seen in the examples in Table 6, which display explicit language in the first four instances, which contain racism. Here, the model can easily detect hate, even though users might even attempt to auto-censor themselves by substituting letters with numbers, as in example #1 (most likely in order to bypass automatic forum filters). Conversely, the misogyny in the last two samples is much more implicit, with the model failing to detect misogyny in sample #5.

Table 6
Examples of racist and misogynous posts from IFS-IT, with gold annotation and Incel mBERT prediction labels.

Post	Gold	Pred.
Compagno le n3gr3 sono oggettivamente brutte, le asiatiche lo sono in media - ma quelle belle lo sono davvero e staccano di misura le cosiddette belle nostrane.	Both	Both
No perchè sessualmente mi fanno schifo le negre e i trans (più questi ultimi eh).	Both	Both
Ora capisco perché non scopava, curry percepito anche se non è curry, currycel in pratica.	Rac.	Rac.
Che dire allora dei terroni quasi tutti arabi quindi negri e di quei rari bianchi europei che provengono da altre nazioni europee? La mafia cioè i terroni stanno importando queste merde in massa per farci terronizzare e negrizzare come loro.	Rac.	Rac.
le 8+ sono davvero rare. tuttavia, dal 5 in su si atteggiano tutte come fossero modelle...	Mis.	None
Probabilmente la vedrai tra qualche settimana ad ipergamare con qualche architetto	Mis.	Mis.

6. Conclusions

In this paper, we have presented an approach to improve the performance of hate speech detection models in Italian incel posts. Our experiments show that domain-adapting transformer models to the contents of incel forums boosts their performance when predicting the hatefulness of incel forum posts, both when using Italian-only and multilingual models. The increase in performance obtained through MLM pre-training is particularly high when using bilingual training data with mBERT_{base}, which might indicate that the model is learning about incel hate speech by learning language-agnostic incel concepts. We have also shown that for the base Italian models (AlBERTo and UmBERTo) fine-tuning on combinations of different Italian corpora can lead to a boost in performance. However, this performance boost is nullified after MLM pre-training, which appears to be a more effective strategy for improving the performance of the models. When looking at racism vs. misogyny identification in posts extracted from *Il forum dei brutti*, the former appears to be much easier to detect. This seems due to the fact that racist language is much more explicit than misogynous language in the scrutinized forum, but further research is needed to ascertain such a supposition.

In future work, we plan to experiment with different resources for MLM pre-training, using corpora in different languages, since it seems multilingual models such as mBERT_{base} are capable of learning about hate speech in a language-agnostic way from multiple languages. In addition, with more computational resources, larger corpora and more training epochs could be used to further improve the performance of the models. Lastly, further experiments can be carried out as regards the performance of the scrutinized models on the individual sub-tasks of misogyny and racism identification, respectively.

References

[1] F. Alkomah, X. Ma, A Literature Review of Textual Hate Speech Detection Methods and Datasets, Information 13 (2022). doi:10.3390/info13060273.

[2] H. Van, Mitigating Data Scarcity for Large Language Models, 2023. doi:10.48550/arXiv.2302.01806.

[3] A. Nagle, Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and the Alt-Right, Zero Books, Winchester, Hampshire, UK, 2017. doi:10.5817/PC2018-3-270.

[4] S. Jaki, T. De Smedt, M. Gwóźdź, R. Panchal, A. Rossa, G. De Pauw, Online Hatred of Women in the Incels.me Forum: Linguistic Analysis and Automatic Detection, Journal of Language Aggression and Conflict 7 (2019) 240–268. doi:10.1075/jlac.00026.jak.

[5] T. Farrell, M. Fernandez, J. Novotny, H. Alani, Exploring Misogyny Across the Manosphere in Reddit, in: Proceedings of the 10th ACM Conference on Web Science, ACM, Boston, MA, 2019, pp. 87–96. doi:10.1145/3292522.3326045.

[6] K. C. Gothard, Exploring Incel Language and Subreddit Activity on Reddit, Honors college senior thesis, UVM Honors College, 2020. URL: https://scholarworks.uvm.edu/hcoltheses/408.

[7] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, Overview of the EVALITA 2018 Hate Speech Detection Task, in: [25], 2018, pp. 67–74.

[8] A. Cimino, L. De Mattei, F. Dell'Orletta, Multi-task Learning in Deep Neural Networks at EVALITA 2018, in: [25], 2018, pp. 86–95.

[9] V. Basile, D. M. Maria, C. Danilo, L. C. Passaro, et al., EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: [26], 2020, pp. 1–7.

[10] E. Lavergne, R. Saini, G. Kovács, K. Murphy, TheNorth @ HaSpeeDe 2: BERT-based Language Model Fine-tuning for Italian Hate Speech Detection, in: [26], 2020.

[11] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Association for Computational Linguistics, Minneapolis, MN, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[12] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR-WS, Bari, Italy, 2019.

[13] L. Parisi, S. Francia, P. Magnani, UmBERTo: An Italian Language Model Trained with Whole Word Masking, https://github.com/musixmatchresearch/umberto, 2020.

[14] E. Fersini, D. Nozza, P. Rosso, Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI), in: [25], 2018.

[15] E. Fersini, D. Nozza, P. Rosso, AMI @ EVALITA 2020: Automatic Misogyny Identification, in: [26], 2020.

[16] A. Muti, A. Barrón-Cedeño, UniBO @ AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo, in: [26], 2020.

[17] C. Di Bonaventura, A. Muti, M. A. Stranisci, O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in Italian, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023), volume 3473, CEUR-WS, Parma, Italy, 2023.

[18] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, in: Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, Montreal, Canada, 2017, pp. 512–515.

[19] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) pp. 14867–14875. doi:10.1609/aaai.v35i17.17745.

[20] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for Abusive Language Detection in English, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Online event, 2021, pp. 17–25. doi:10.18653/v1/2021.woah-1.3.

[21] A. Pelicon, R. Shekhar, B. Škrlj, M. Purver, S. Pollak, Investigating Cross-Lingual Training for Offensive Language Detection, PeerJ Computer Science 7 (2021) e559. doi:10.7717/peerj-cs.559.

[22] O. Gokhale, A. Kane, S. Patankar, T. Chavan, R. Joshi, Spread Love Not Hate: Undermining the Importance of Hateful Pre-training for Hate Speech

Detection, 2022. URL: http://arxiv.org/abs/2210.04267.

[23] P. Gajo, A. Muti, K. Korre, S. Bernardini, A. Barrón-Cedeño, On the Identification and Forecasting of Hate Speech in Inceldom, in: G. Angelova, M. Kunilovskaya, R. Mitkov (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2023), volume 2263, INCOMA Ltd., Varna, Bulgaria, 2023, pp. 373–384.

[24] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task, in: [26], 2020.

[25] T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), volume 2263, CEUR-WS, Turin, Italy, 2018.

[26] V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), volume 2765, CEUR-WS, Online event, 2020.