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

Paolo Gajo, Alberto Barrón Cedeño, Silvia Bernardini, Adriano Ferraresi
University of Bologna, Italy
paolo.gajo@studio.unibo.it
{a.barron, silvia.bernardini, adriano.ferraresi}@unibo.it

Abstract

English. In this study, we aim to enhance hate speech detection in an Italian incel forum. 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, binary-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 un forum italiano di incel. Addestriamo modelli Transformer monolingue (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 binarie 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.*

1 Introduction

Hate speech, broadly defined as language that expresses hatred towards a targeted group or is intended to be derogatory, humiliating, or insulting to the members of the group (Davidson et al., 2017), has become an increasingly prevalent and dangerous phenomenon (Matamoros-Fernández and Farkas, 2021). A specific area of concern in the realm of hate speech is the online spaces known as the "Manosphere", where misogynous discourse in particular has become increasingly rampant (Ribeiro et al., 2021). Specifically, within the Manosphere, the incel (short for "involuntary celibate") community has been identified as frequently engaging in hateful, misogynous, and racist speech (Nagle, 2017; Jaki et al., 2019).

While there is no scarcity of English-language models and training resources for the detection of hate speech, especially with the recent rise in popularity of this research topic (Alkomah and Ma, 2022), much work can still be done when approaching 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 (Van, 2023). The problem becomes even more exacerbated when considering the detection of hate speech in niche contexts, such as in forums frequented by incels, which are characterized by the use of specific misogynous and racist lexicon (Gothard, 2020). 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 improving the performance of hate speech detection models in Italian incel forums. Our contribution is two-fold:

(*i*) Masked language modeling We adapt monolingual Italian models and multilingual models to the linguistic domain of Italian and English incel forums by training them on the masked language modeling (MLM) task. As training material, we use unsupervised corpora built from two incel forums, one in Italian and one in English. We release these novel models, which can be used for further research on the topic.[1]

***(ii)* Hate speech detection** We approach the detection of hate speech in Italian incel forums using monolingual (Italian) and bilingual (Italian–English) combinations of supervised corpora. The corpora were compiled from incel forums and mainstream social media platforms and are binary-annotated for hate speech. We fine-tune the monolingual models on Italian-only combinations of these corpora, while the multilingual models are fine-tuned on bilingual combinations. Testing their performance on a supervised hate speech corpus, obtained from the aforementioned Italian incel forum, shows that the best results are obtained by training the 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, binary-annotated for hate speech.

In the approached scenarios, pre-training and fine-tuning on in-domain incel annotated data may therefore be more effective than training on general target-language supervised corpora, despite part of the training data not being in the language of the downstream 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 models used in this study, while Section 5 describes the experiments conducted and discusses the results.

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 (Bosco et al., 2018) based on two corpora, one from Twitter and one from Facebook. The participating teams experimented with a variety of machine-learning and deep-learning algorithms, with the top team relying on an SVM and a BiLSTM (Cimino et al., 2018). The 2020 edition hosted a shared task on the detection of hate speech, especially against migrants, focusing on tweets and news headlines (Basile et al., 2020). In this case, the best results were obtained by using AlBERTo (Polignano et al., 2019) and UmBERTo (Parisi et al., 2020), two BERT models pre-trained on Italian data. The 2020 edition also hosted a shared task on the automatic identification of misogyny in Italian tweets (Fersini et al., 2020), where an ensemble of BERT models won the competition (Muti and Barrón-Cedeño, 2020).

English-language hate speech detection has been conducted in a variety of ways. Davidson et al. (2017) 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. (2021) build a dataset called "HateXplain" from Twitter and Gab posts, annotated with a multi-class label based on whether the post is "offensive", expesses "hate", or is "normal", which they use to fine-tune a BERT hate speech classifier. Caselli et al. (2021) retrain BERT$_{base}$ on the MLM task using a dataset 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. (2021) use a multilingual combination of datasets annotated for hate speech to improve the performance of classifiers in zero-shot, few-shot and well-resourced settings. Gokhale et al. (2022) use MLM training to improve the hate speech detection performance of BERT in Hindi and Marathi, separately. We follow such approaches in attempting to improve the performance of our models.

3 Corpora

We leverage three supervised Italian-language corpora from past EVALITA campaigns, along with two supervised corpora compiled from two incel forums, one in English and one in Italian.

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 (Bosco et al., 2018) (henceforth "HSD-FB"), by annotating Facebook posts for hate speech. The second one is from the

[1] https://github.com/paolo-gajo/clic23

Table 1: Statistics of the *Incels.is* and *Il forum dei brutti* (*FdB*) unsupervised corpora. Mean length computed at token level.

Corpus	Posts	Threads	Length
Incels.is	4,760k	230k	31.07±70.01
FdB	638k	30k	52.78±80.77

2020 HaSpeeDe shared task (Basile et al., 2020) ("HSD-TW"), compiled by adding new data to the HaSpeeDe 2018 Twitter dataset. The third corpus is the one compiled for the Automatic Misogyny Identification (AMI) shared task (Fersini et al., 2020) ("AMI-20"), hosted at EVALITA 2020. AMI-20 also contains tweets and is annotated with misogyny labels, which we use in place of hate speech labels.[2] The corpora were split 70/30 between training and development sets (we do not use the test partitions of these shared tasks).

Incel corpora We use two unsupervised corpora compiled by scraping two incel forums: *Incels.is*[3] and *Il forum dei brutti*[4] (Gajo et al., 2023). Table 1 reports the statistics of the two corpora.[5] We built these new resources both due to the lack of freely available incel corpora and the fact that incel language changes rapidly, making it worthwhile to compile updated resources.

P: do i need to cite myself again here for the claim that incel language changes rapidly?

A subset of the two datasets was annotated for both misogyny and racism. We refer to these two supervised corpora as IFS-EN (Incel Forum, Supervised, English) and IFS-IT (Incel Forum, Supervised, Italian). IFS-EN is split between training, development and testing partitions with a 70/15/15 split. In this study, IFS-IT is only used as a test set, meaning it is unpartitioned and all of its 500 instances are used for testing. The datasets are annotated for racism and misogyny because they are the most relevant forms of hate speech in the two forums (Silva et al., 2016; Ging and Siapera, 2018). However, it has to be noted that racism is much more prevalent in *Incels.is* than in *Il forum dei brutti*. Posts are annotated as hateful if they are either labeled as misogynous or racist.

Table 2 shows the class distribution of all five supervised corpora.

Table 2: Class distribution for the different supervised hate speech corpora.

Corpus		HS	Non-HS
AMI-20	(Fersini et al., 2020)	2,337	2,663
HSD-FB	(Bosco et al., 2018)	1,382	1,617
HSD-TW	(Basile et al., 2020)	971	2,028
IFS-EN	(Gajo et al., 2023)	2,090	3,113
IFS-IT	(Gajo et al., 2023)	200	300

4 Models

With relation to the Italian-only scenario, we use UmBERTo and AlBERTo as our baseline models. We choose these models because they achieved the best performance in previous EVALITA shared tasks on hate speech (Basile et al., 2020) and misogyny (Fersini et al., 2020) 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 (Caselli et al., 2021) and incel forums (Gajo et al., 2023). For training data, we use the entirety of the contents of the forum, for a total of 627k posts.[6] In doing this, we obtain two new models which we refer to as "Incel UmBERTo" and "Incel AlBERTo".

P: should I say that we do this also bilingually based on the ranlp paper? otherwise it seems kind of strange that i just randomly decided to do bilingual MLM and on top of that we get the best results with the bilingual model

As regards the bilingual setting, we use mBERT$_{base}$ as our baseline. We also use an MLM-enhanced version of it, "Incel mBERT", which we obtain by training it for one epoch 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.

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 a single Tesla P100 GPU with 16 GB of VRAM.

5 Experiments and Evaluation

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 previously introduced datasets. We do this in order to make our results more reliable and diminish the ef-

[2] In incel spaces most of the hate speech is expressed in the form of misogyny, making this data useful.

[3] https://incels.is

[4] https://ilforumdeibrutti.forumfree.it

[5] Both corpora are available at: https://zenodo.org/record/8147845.

[6] The number of posts is lower than the total number of posts (638k) because we exclude empty posts.

Table 3: Performance when fine-tuning on Italian-only corpora combinations. Epochs (e) selected based on validation performance. Highest scores in bold.

	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

fect 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. For each dataset 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 score 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 (Bosco et al., 2018; Basile et al., 2020) and

Table 4: Performance when fine-tuning on monolingual and bilingual combinations of English and Italian supervised corpora. Epochs (e) selected based on validation performance. Highest scores in bold.

		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

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" (Bosco et al., 2018). The fact that messages in HSD-FB are longer is also coherent with the Italian incel models performing better than the vanilla models when training on HSD-FB, since *Il forum dei brutti* on average contains rather long posts (see Table 1), unlike Twitter datasets, 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 datasets, 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 datasets. Therefore, while some improvement can be observed by merging different datasets, 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$_1$ score of 0.688, the best combination using Incel mBERT achieves a test F$_1$ score 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 AMI-20 dataset 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 we are using posts extracted from this forum to pre-train Incel mBERT, 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.

On average, the lowest performance is achieved when training separately on IFS-EN and the Italian corpora. When using bilingual data, the worst results are obtained when training on HSD-TW and the combinations containing it, which is coherent with the results obtained in the monolingual setting. For almost all combinations of Italian corpora, performance increases once IFS-EN is added to the training data, i.e., bilingual combinations lead 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 dataset combinations, but on the bilingual ones. Indeed, for four bilingual dataset 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 surprising. Even more interesting is the fact that, although we are testing on an entirely Italian dataset, 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 seems to yield 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 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 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.

6 Conclusions

In this paper, we have presented an approach to improve the performance of hate speech detection models within Italian incel forums. 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 datasets 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.

In future work, we plan to experiment with different resources for MLM-pretraining, 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, more training epochs could be used to further improve the performance of the models.

References

Fatimah Alkomah and Xiaogang Ma. 2022. A Literature Review of Textual Hate Speech Detection Methods and Datasets. *Information*, 13(6):273, May.

Valerio Basile, Di Maro Maria, Croce Danilo, Lucia C Passaro, et al. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and sspeech tools for italian. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, pages 1–7. CEUR-ws.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian*, pages 67–74. Accademia University Press.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English, February.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task learning in deep neural networks at evalita 2018. *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, pages 86–95.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*. arXiv, May.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic Misogyny Identification. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 21–28. Accademia University Press.

Paolo Gajo, Arianna Muti, Katerina Korre, Silvia Bernardini, and Alberto Barrón-Cedeño. 2023. On the identification and forecasting of hate speech in inceldom. In *Proceedings of the 2023 International Conference on Recent Advances in Natural Language Processing (RANLP 2023)*. To appear.

Debbie Ging and Eugenia Siapera. 2018. Special issue on online misogyny. *Feminist Media Studies*, 18(4):515–524.

Omkar Gokhale, Aditya Kane, Shantanu Patankar, Tanmay Chavan, and Raviraj Joshi. 2022. Spread Love Not Hate: Undermining the Importance of Hateful Pre-training for Hate Speech Detection. December. arXiv:2210.04267 [cs].

Kelly C Gothard. 2020. Exploring Incel Language and Subreddit Activity on Reddit.

Sylvia Jaki, Tom De Smedt, Maja Gwóźdź, Rudresh Panchal, Alexander Rossa, and Guy De Pauw. 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, November.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media*, 22(2):205–224, February.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Arianna Muti and Alberto Barrón-Cedeño. 2020. UniBO @ AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 29–34. Accademia University Press.

Angela Nagle. 2017. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right.* Zero Books, Winchester, Hampshire, UK, July.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. `https://github.com/musixmatchresearch/umberto`.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559, June.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 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. CEUR.

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2021. The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 196–207.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Tenth international AAAI conference on web and social media.*

Hoang Van. 2023. Mitigating data scarcity for large language models.