Hate Speech Detection in Italian Incel Forums Using Bilingual Fine-Tuning Data

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 paper we approach the problem of improving the performance of hate speech detection models in cross-lingual and cross-domain settings, in Internet forums frequented by "incels", short for "involuntary celibates". We pre-train Transformer models on corpora built from two incel forums, one in Italian and one in English, using masked language modeling. We fine-tune the models on English and Italian datasets annotated for hate speech. Testing the performance of the fine-tuned models on a hate speech dataset obtained from an Italian incel forum shows that the best results are achieved by using a combination of Italian and English data, rather than just Italian, emphasizing the importance of in-domain data, even when languages differ.

Italiano. *In questo studio affrontiamo il problema dell'identificazione dei discorsi d'odio in forum frequentati da "incel", abbreviazione inglese di "celibi involontari". Addestriamo quindi modelli Transformer su corpus costruiti a partire da due forum di incel, uno in italiano e uno in inglese, con il masked language modeling. Quindi, facciamo il fine-tuning dei modelli su dataset in italiano e inglese con annotazioni indicanti se un post esprime discorsi d'odio. Valutando le prestazioni dei modelli su un dataset ottenuto da un forum italiano di incel, osserviamo che i risultati migliori si ottengono addestrandoli su una combinazione di dataset italiani e inglesi, piuttosto che solo italiani, sottolineando l'importanza di utilizzare dati appartenenti allo stesso contesto linguistico, anche quando le lingue differiscono.*

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 in the past years (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, the incel (short for "involuntary celibate") community within the Manosphere has been identified as one that frequently engages 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 cross-lingual and cross-domain settings, focusing on detection efforts on Italian incel forums. Our contributions are the following:

(*i*) Corpora. We compile two novel unsupervised corpora on the domain of incel forums, one in Ital-

ian and one in English. We annotate a subset of each for hate speech, misogyny, and racism, obtaining two supervised corpora. The unsupervised corpora can be used for domain-adaptation, while the supervised ones can be used for training hate speech detection models.

***(ii)* Masked language modeling.** For the first time, 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 on the aforementioned unsupervised corpora. We release these novel models, which can be used for further research on the topic.[1]

***(iii)* Hate speech identification.** We approach the identification of hate speech in Italian incel forums using monolingual (Italian) and bilingual (English-Italian) training data. We train the obtained MLM-enhanced models on various combinations of English and Italian datasets, pertaining to the domains of incel forums and mainstream social media, testing. Then, we test their hate speech identification performance on the domain of Italian incel forums.

Our experiments show that the best results are obtained by training the mBERT model on bilingual data taken from both the English and Italian incel forums, using the MLM task, and then fine-tuning it on combinations of English and Italian datasets annotated for hate speech. This shows how in this niche scenario having in-domain incel annotated data may be more effective than using general target-language instances, despite part of the training data not being in the target language.

2 Related Work

Prior work on Italian hate speech identification has been conducted chiefly within the context of the EVALITA shared tasks. The 2018 edition hosted a shared task on hate speech detection (Bosco et al., 2018) based on two Italian-language datasets annotated for hate speech, 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, based on a dataset of 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 Italian-language Transformer models. 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 achieved the top performance (Muti and Barrón-Cedeño, 2020).

English-language hate speech detection has been conducted both with models such as logistic regression and linear SVMs (Davidson et al., 2017) and, more recently, by using Transformers (Mathew et al., 2021).

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 Datasets

We leverage existing Italian-language datasets from past EVALITA campaigns, along with two datasets compiled from two incel forums, one in English and one in Italian. All of the dataset are binary-annotated at a post level for hate speech.

Existing Italian datasets. The first Italian dataset we use was compiled for the first edition of the Hate Speech Detection (HaSpeeDe) shared task, hosted at EVALITA 2018 (Bosco et al., 2018) (henceforth "HSD-FB"), by annotating Facebook posts for hate speech. The second one is from the 2020 HaSpeeDe shared task (Basile et al., 2020) ("HSD-TW"), compiled by adding new data to the HaSpeeDe 2018 Twitter dataset. The third and last dataset we use in Italian is the one compiled for the Automatic Misogyny Identification (AMI) shared task (Fersini et al., 2020) ("AMI-20"), hosted at EVALITA 2020. This dataset is also compiled from tweets and is annotated with misogyny labels, which we use in place of hate speech labels.[2] All Italian datasets were split 70/30 between training and development sets. We do not use the test sets of these shared tasks.

Novel incel datasets. We compiled two novel

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

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

Table 1: Statistics of the IFC-22-EN and IFC-22-IT unsupervised datasets in terms of posts and threads. Mean length computed at token level.

Dataset	Posts	Threads	Length
IFU-22-EN	4,760k	230k	31.07±70.01
IFU-22-IT	638k	30k	52.78±80.77

Figure 1: Guidelines for the corpus annotation, derived from (Fersini et al., 2018) for misogyny and (Waseem and Hovy, 2016) for racism.

Please identify whether each post is categorized as misogynous, racist, or falls into another category:
A post is deemed **misogynous** if it exhibits any of the following traits:

- Objectifies or stereotypes women;
- Claims that men are superior to women;
- Derails the conversation to defend the abuse of women, deny male responsibility, or redirect the conversation in favor of men;
- Contains sexual advances, solicits sexual favors, sexually harasses the recipient, or threatens women with physical violence to assert power;
- Uses slurs against women without any legitimate purpose.

A post is considered **racist** if it exhibits any of the following traits:

- Uses a racial slur;
- Stereotypes, attacks, or seeks to silence a minority without a valid argument;
- Promotes violent crime against minorities;
- Misrepresents the truth or distorts views on a minority with baseless claims;
- Shows support for problematic ideologies, such as xenophobia, homophobia, or sexism.

unsupervised datasets, comprising posts scraped from incel forums: IFU-22-EN (Incel Forum, Unsupervised, 2022, English) scraped from the *Incels.is* forum[3], and IFU-22-IT (Incel Forum, Unsupervised, 2022, Italian) scraped from *Il forum dei brutti*[4]. The posts were scraped making sure to retain all metadata associated with them, with the main content being saved separately from the quoted content which a user is replying to. Table 1 reports the statistics of the two datasets.[5] We build these new resources both due to the lack of freely available incel corpora and the fact that incel language changes rapidly, as outlined in Appendix A, making it worthwhile to compile updated resources.

[3] https://incels.is

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

[5] The datasets are available at: https://zenodo.org/record/8147845.

Table 2: Hate speech (HS) annotation statistics for the adopted datasets.

Dataset	HS	%	Non-HS	%
IFS-EN	2,090	40.17	3,113	59.83
IFS-IT	200	40.00	300	60.00
HSD-FB	1,382	46.08	1,617	53.92
HSD-TW	971	32.38	2,028	67.62
AMI-20	2,337	46.74	2,663	53.26

A subset of the two datasets was annotated for both misogyny and racism, following the guidelines presented in Figure 1. For English, a subset of 50 posts was randomly selected and labeled by three annotators, all with a C2 CEFR level of English, well-versed in linguistics, gender studies, NLP, and data annotation. The Cohen's Kappa inter-annotator agreement (IAA) (Bobicev and Sokolova, 2017) was of 0.77, which is considered high. As such, the remaining instances were annotated by a single annotator. For the Italian dataset, two native speakers of Italian, also experts in the relevant fields, annotated 50 posts as well, reaching an IAA of 0.69. As the IAA was deemed acceptable, the rest of the instances were once again annotated by one annotator.

We refer to these two supervised datasets 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, while IFS-IT is only used for testing. The datasets are annotated for racism and misogyny, since they are the most relevant forms of hate speech in the two forums (Silva et al., 2016; Ging and Siapera, 2018). Posts are annotated as hateful if they are either labeled as misogynous or racist.

The annotation statistics of all the datasets used in this study can be found in Table 2.

4 Models

We train the Italian-only models on combinations of the Italian datasets and the multilingual models on English-Italian combinations.

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. We also create MLM-enhanced versions of these two models by training them on the en-

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

	IT			(e)	Validation (IT)			Test (IFS-IT)		
	HSD-FB	HSD-TW	AMI-20		F_1	Rec	Prec	F_1	Rec	Prec
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

tirety of the contents of IFU-22-IT, for a total of $627k$ sentences.[6] We refer to these models as "Incel UmBERTo" and "Incel AlBERTo".

As regards the Italian-English setting, we use mBERT$_{base}$ as our baseline. We also use an MLM-enhanced version of it which we obtain by training it for one epoch on $500k$ posts sampled from IFU-22-EN and $500k$ posts sampled from IFU-22-IT, for a total of $1M$ bilingual sentences. We refer to this model as "Incel mBERT".

The MLM pre-training process is carried out in all cases by tokenizing the sentences using Hugging Face's AutoTokenizer[7], feeding them into the model using Hugging Face's data collator for language modeling.[8] Due to hardware constraints, we train the models only for one epoch, with a masking probability of 15% and a batch size of 32 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. 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 perfor-

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

[7]https://huggingface.co/docs/transformers/model_doc/auto

[8]https://huggingface.co/docs/transformers/main/main_classes/data_collator

Table 4: Performance when fine-tuning on Italian-English dataset combinations. Epochs (e) are selected based on the convergence of the validation performance. Highest scores in bold.

	IT			EN	(e)	Validation (IT-EN)			Test (IFS-IT)		
	HSD-FB	HSD-TW	AMI-20	IFS-EN		F_1	Rec	Prec	F_1	Rec	Prec
mBERT	■			■	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	■			■	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

mance on the validation set. We do this in order to make our results more reliable and diminishing the effect of the random initialization of the models. For each dataset combination, the training and validation sets are the unions of the individual training and validation sets of each merged dataset. The resulting models are then evaluated on the IFS-IT test set.

Table 3 shows the performance in terms of precision, recall and F_1-measure for the Italian-only models and dataset combinations, while Table 4 reports the results for the bilingual setting.

Monolingual setting. Overall, the best performing model on the Italian-only combinations is Incel AlBERTo, which achieves a test F_1 score of 0.707, when training solely on HSD-FB. Compared to training AlBERTo on the same dataset, which achieves a test F_1 score of 0.683 on IFS-IT, this represents a performance boost of 2.4 F_1 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 pretraining AlBERTo and UmBERTo on Italian sentences extracted from IFU-22-IT is effective in improving their performance.

The worst test results, for all models, 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 noticeably lower for HSD-TW combinations, compared to HSD-FB combinations. This shows that the models have a harder time learning from HSD-TW than from HSD-FB. 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 with the fact that the messages contained in HSD-FB 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 the messages contained in HSD-FB are longer would also be coherent with the fact that Incel AlBERTo and Incel UmBERTo perform better than UmBERTo and AlBERTo when training on HSD-FB, since IFU-22-IT on average contains rather long posts. Finally, an additional element which might explain the lower performance when training on HSD-TW is the fact that it contains hate speech against migrants, which might not be as relevant when it comes to identifying hate speech in incel spaces.

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 IFU-22-IT, the performance boost is no longer observed. This could be due to

a high affinity between IFU-22-IT and HSD-FB, compared to HSD-TW and AMI-20, which would mean that pre-training on IFU-22-IT biases the model toward the language of that dataset, reducing performance when adding others to the mix.

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 Italian-only 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. 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-ENare misogynous), and due to the fact that we are using sentences extracted from this forum to pre-train Incel mBERT, the performance boost could be due to the fact that the model is better at detecting misogynous language compared to mBERT$_{base}$ and the Italian-only models.

The lowest performance is once again achieved when only training on HSD-TW (and IFS-EN). The combinations containing this dataset also lead to lower performance, both for mBERT$_{base}$ and Incel mBERT. For example, the performance difference between training Incel mBERT only on AMI-20 and HSD-TW+AMI-20 is 2.2 F_1 points in favor of the former combination.

IT-EN vs. IT. The results of our experiments show that the highest performance is not obtained by training on the Italian-only datasets, but rather on the English-Italian ones. For five dataset combinations out of seven, Incel mBERT's performance is higher than all other models. The only combinations for which the performance of Incel mBERT does not beat all other models are HSD-FB+AMI-20 and HSD-TW+AMI-20.

Since mBERT was originally pre-trained on 104 languages and AlBERTo and UmBERTo were pre-trained only on Italian, 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. This shows that, in this case, using bilingual instances on a multilingual model for MLM may result in higher performance than pre-training Italian models only on Italian sentences. Furthermore, the number of Italian sentences used to train Incel AlBERTo and Incel UmBERTo is $627k$, which is greater than the $500k$ Italian sentences used for Incel mBERT.

From these results, we could infer that the model is learning more about how to spot hate speech in IFS-IT by learning language-agnostic incel concepts, since Incel mBERT is pre-trained on sentences extracted from two incel forums in two different languages. Although the two speech communities are obviously distinct, the red pill and black pill ideologies have spread internationally, and are shared by both communities. 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.

6 Conclusions

In this paper, we have presented an approach to improve the performance of hate speech detection models in a cross-lingual and cross-domain scenarios. We used existing datasets in Italian annotated for hate speech and combined them with unsupervised datasets compiled by us in English and Italian from incel forums.

Choosing mBERT as our baseline model, we improved its performance by following the approach of Caselli et al. (2021) through MLM training, enhancing its capability of detecting hateful language in the target language. We did the same for Italian models, namely UmBERTo and AlBERTo, and compared the results with the baseline mBERT models.

Our experiments show that

Future work

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.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *International Conference Recent Advances in Natural Language Processing*, pages 97–102.

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, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. 2150:214–228.

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.

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.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The tenten corpus family. *7th International Corpus Linguistics Conference CL 2013*, 07.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*.

Lexical Computing Ltd. 2015. Statistic used in sketch engine. `https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/`, 7.

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.

Björn Pelzer, Lisa Kaati, Katie Cohen, and Johan Fernquist. 2021. Toxic language in online incel communities. *SN Social Sciences*, 1(8):1–22.

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.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. 2022. Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering*, pages 1–28.

A Analysis of Keyness in Incel Forums

We can investigate the difference of relative frequency in word usage between general language and the language used in a specific speech community by building corpora representative of the two groups of speakers. That is, we can use a large *reference corpus*, representing general language usage, and compare its frequencies to a *focus corpus* (Kilgarriff, 2009), built only from texts pertaining to a specific communicative context.

We show the evolution of incel language by studying the change in *keyness* (Kilgarriff, 2009) of specific sets of words, showing how the lexical features of incel communities of speakers of English and Italian change rapidly over time. Keyness indicates which words in a focus corpus are highly frequent compared to a reference corpus. The keyness of a word w is defined as (Lexical Computing Ltd., 2015):

$$keyness(w) = \frac{fpm_f(w) + n}{fpm_r(w) + n} \qquad (1)$$

where $fpm_f(w)$ represents the normalized frequency of a focus corpus word per million words, $fpm_r(w)$ refers to the word in the reference corpus, and n is a smoothing parameter (here, $n = 1$).

To study the English-speaking *Incels.is* forum, we consider all of its contents, for a total of $104M$ words (collected up to 18 October 2022). We do the same for the Italian *Il forum dei brutti*, for a total of $30M$ words (up to 4 December 2022). For English, we calculate the keyness by using enTenTen20 as the reference corpus, while for Italian we use itTenTen20 (Jakubíček et al., 2013).

As regards *Incels.is*, in order to compile a list of characteristic incel lexicon, the keyness of lexical items was calculated across the entirety of the forum, up to October 2022. Preliminary candidates were selected by collecting single- and multi-word items that ranked in the top 500 for keyness, for a total of $1k$ analyzed items. Among these, only terms considered to be typical of incel language were examined. R'acism and misogyny are very characteristic elements of the language of incels. As such, a simple way to choose characteristic terms for this speech community is manually evaluating racist and misogynous terms (or terms that are frequently associated to racist and misogynous contexts) and selecting those which are not typically found in general language, i.e., having high keyness scores. The evaluation of the individual terms was carried out by manually analyzing concordance lines in the corpus with the objective of verifying whether their use could be construed as being hateful. Although human evaluation is unavoidably subjective, we erred on the side of caution and only selected terms which could unmistakably be used in a hateful manner. Unfortunately, this terminology extraction strategy has the drawback of not directly taking into account terms that get resemanticized and assume a new, offensive meaning. Further work could be carried out to identify such terms in order to have a more comprehensive understanding of the issue.

With relation to *Il forum dei brutti*, we once again studied terms we deemed to be characteristic of the forum's incel language; however, in this case we focused on 10 terms used to describe other men in negative or positive ways. We chose this approach because the goal of this modern diachronic study is to show that language specific to incels changes over time, regardless of whether it can be considered hateful. Therefore, since in IFU-22-IT we could not find as much misogynous or racist jargon as in IFU-22-EN, we decided to consider the way men are represented, instead of women.

In order to conduct the study, the contents posted on the *Incels.is* forum from 2017 to 2022 were divided into 22 chronological partitions, one for each 100 pages, each page containing 100 threads. With a similar approach, *Il forum dei*

brutti was divided chronologically by grouping posts by year of creation, from 2009 to 2022, for a total of 14 partitions.

The keyness of each selected term was measured for every partition, calculating the slope m of its regression line as:

$$m = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(k_i - \bar{k})}{\sum_{i=1}^{n}(t_i - \bar{t})^2} \quad (2)$$

where t_i is the i-th time partition, k_i is the i-th keyness score, n is the number of partitions, and $\bar{t}$ and $\bar{k}$ are the means of the two variables. By calculating the slope of the regression line, we are able to find how the keyness of a term changes over time. A positive slope indicates that the use of a term is becoming more frequent, while a negative slope indicates that a term is becoming less prevalent. For each term, the slope was first calculated across all partitions (22 for *Incels.is* and 14 for *Il forum dei brutti*); then, it was divided by the average keyness of the term over all the partitions, thus obtaining the normalized slope. This was done because certain terms may have very high keyness values, while other terms may not be as prevalent, and we wanted to be able to compare the slope of different terms regardless of the absolute value of their keyness.

For each partition, only the keyness of the 500 terms with the highest keyness was recorded. Zero values, produced whenever the item's keyness was not high enough to appear among the top 500 terms of the partition, were ignored both for the calculation of the slope and the average keyness. The number of zero values for *Incels.is* was 7.16% of the total, while for *Il forum dei brutti* it was 44.44%.

With relation to *Incels.is*, we selected the 10 terms with the highest and lowest normalized slope, 20 in total, while for *Il forum dei brutti* we only picked the top and bottom 5 terms, 10 in total. The lower number of terms for *Il forum dei brutti* is due to the fact that we could not identify enough relevant terms for the study. For both forums, the mean normalized slope was finally calculated for each group of terms to have a pair of values which could be used to compare the two overall trends.

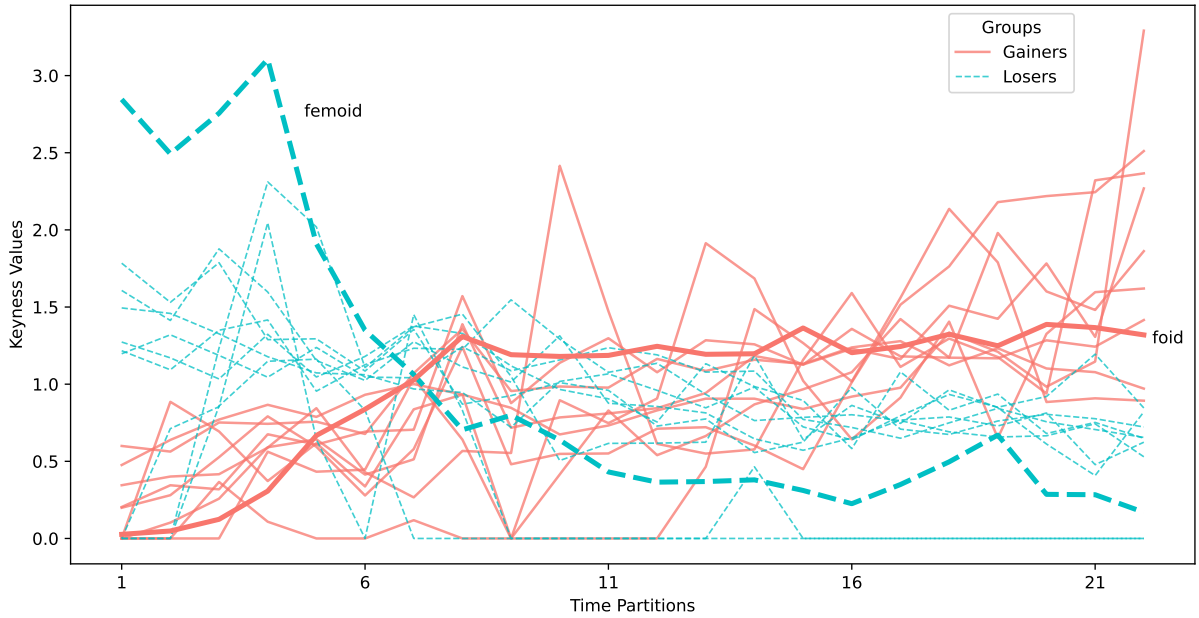
Figure 2 shows the over-time trend of the keyness of the terms extracted from *Incels.is* and *Il forum dei brutti* over the partitions of the two forums. The curves show clear opposite trends for the two groups, which we refer to as "gainers" and "losers" of keyness, based on whether their mean normalized slope is positive or negative, respectively. The plots help visualize a widening over-time difference in lexicon, which may cause models trained on dated texts to become increasingly worse at evaluating more recent data. The highlighted terms in the figure also show that certain terms seem to substitute each other over time, although not all of them can be paired in this manner. For example, "foid" is a contraction of "femoid" and "adone" is a close synonym of "chad", and for both pairs we can observe opposite trends with a specific point in time in which one overtakes the other.

Table 5: Keyness normalized slopes for *Incels.is* and *Il forum dei brutti*.

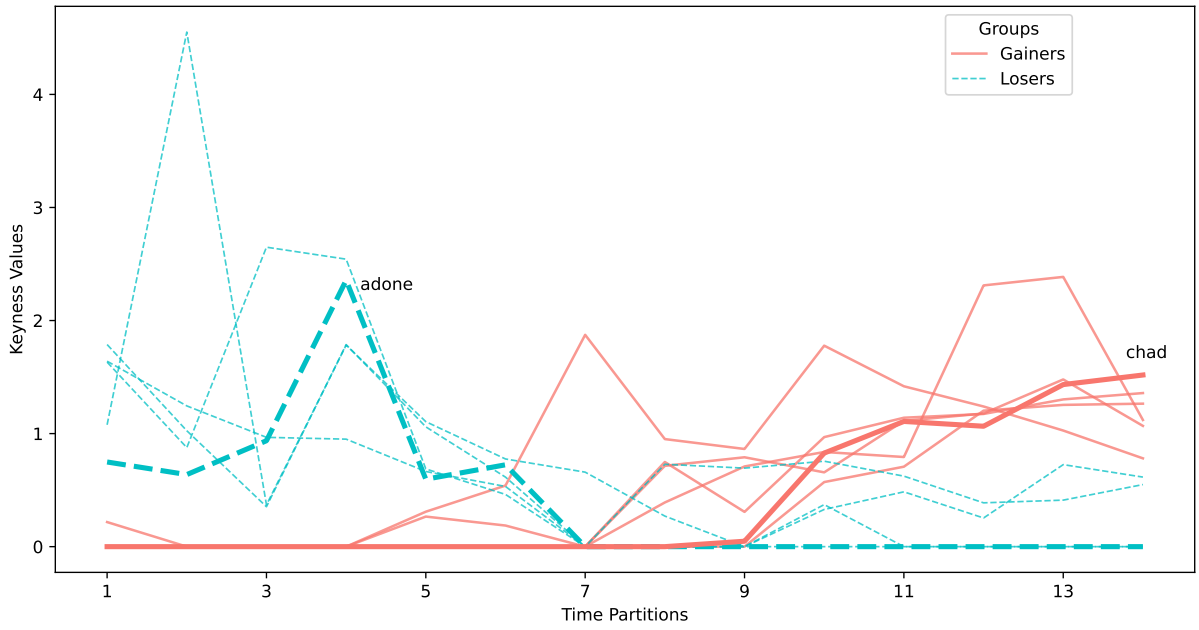
Forum	Gainers		Losers	
	Term	**Slope**	**Term**	**Slope**
Incels.is	shitskin	0.093	racepill	-0.019
	deathnic	0.081	stacie	-0.022
	cumskin	0.079	jb	-0.027
	noodlewhore	0.077	chadlite	-0.029
	slav	0.068	whitecels	-0.032
	foid	0.058	cunt	-0.036
	curryland	0.051	slut	-0.046
	aryan	0.048	deathnik	-0.047
	ricecel	0.047	roastie	-0.051
	whore	0.025	femoid	-0.124
	Mean	**0.063**	**Mean**	**-0.043**
Il forum dei brutti	zerbini	0.104	reietto	-0.142
	normie	0.121	strafigo	-0.122
	bv	0.125	figaccione	-0.122
	chad	0.126	attraente	-0.113
	subumano	0.158	adone	-0.103
	Mean	**0.127**	**Mean**	**-0.120**

Table 5 reports the normalized slopes of the terms obtained from the two forums. In both cases, the mean normalized slopes of the two data series, compared side by side, quantitatively display a clear trend according to which certain terms gain popularity over time, while others become less popular. With regard to *Incels.is*, the difference between the mean normalized slopes is 0.106, while for *Il forum dei brutti* the difference is even larger, 0.247, which points at an even faster lexical evolution. In both cases, the shift in lexicon needs to be taken into account in order to have a clear picture of the language adopted by each speech community. For terms such as "foid", "femoid", and "roastie", the observed trends also confirm the time-series data discussed in Gothard (2020), which show certain terms increasing and decreasing in use over the total messages posted in incel subreddits.

With relation to *Incels.is*, as already anticipated



(a) *Incels.is*



(b) *Il forum dei brutti*

Figure 2: Keyness over time for the characteristic incel terms extracted from the (a) *Incels.is* and (b) *Il forum dei brutti* forums. Red lines represent the terms that gained keyness over time, while blue lines represent the terms that lost keyness over time.

through Figure 2, although terms like "foid" and "femoid" have the same meaning (both are used to dehumanize women by associating them to insentient androids[9]), the shorter form has become more popular, while the use of the full form has decreased. This is probably due to the fact that, given the high frequency with which the term is used in the forum, users tend to use the abbreviated version to save time and effort. This might seem like a minor detail, but the sheer amount of misogyny that is expressed in the forum through this term alone makes it important to point out a shift in its use.

As regards *Il forum dei brutti*, we can observe that the way users refer to men changes in a rather clear way. On one hand, positive words that are commonly used in general language, such as "strafigo" and "figaccione" (both meaning "extremely handsome"), are substituted by specialized terms that are more specific to the forum's speech community, e.g., "chad".[10] On the other hand, we can see the same phenomenon for negative words, where "reietto" ("outcast") loses popularity, leaving space to terms with more specialized uses, such as "bv", meaning "brutto vero" (lit. "truly ugly"), and "subumano", meaning "subhuman". The first is an acronym, which makes its meaning opaque to outsiders, while the second is a term with a much stronger and denigrating connotation.

Based on the conducted qualitative and quantitative analyses, the same conclusions can be drawn for both forums: the presented terms are arguably characteristic of the incel language used within the two platforms and the change in their usage over time is non-negligible. This implies that language models could become progressively worse at predicting over these domains, were their training resources not be periodically updated. Models rely on training material to learn language, and if the material is outdated, their understanding of the discourse currently produced by a specific speech community could become suboptimal. This is especially important considering the fact that, especially in the case of *Incels.is*, the presented racist and misogynous terms are novel and carry most of the discriminatory meaning through neologisms.

Consequently, it seems desirable, if not necessary, to periodically update corpora to have accurate terminological representations. In some cases, it would arguably make sense to even rebuild resources from scratch, were they too outdated. In our case, given the observed changes in keyness, we estimate that the hereby analyzed time frame could be taken as a reference for how long resources can be considered up-to-date. However, with the aim of obtaining an objective figure, further research could be conducted to quantify how often resources should be updated to keep up with the evolution of the language used in the spaces scrutinized through this study.

The necessity to build such material is also supported by the fact that resources on the topic of incels are rare and limited, and their applicability is often compromised because the linguistic domain of the source data only partially aligns with the one under investigation (Pelzer et al., 2021). An additional cause for such incompatibility of resources can be found in the annotation scheme, which can be inapplicable to the supervised task being approached (Zhou et al., 2022). However, the necessity to build new resources does not mean they will be obsolete soon after being employed, as the time frames we have analyzed in this chapter span various years of forum activity.

[9] https://incels.wiki/w/Femoid
[10] https://incels.wiki/w/Chad