On the Identification and Forecasting of Hate Speech in Inceldom

Paolo Gajo, Arianna Muti, Katerina Korre,
Silvia Bernardini and **Alberto Barrón-Cedeño**
DIT, Università di Bologna, Forlì, Italy
paolo.gajo@studio.unibo.it
{arianna.muti, aikaterini.korre, silvia.bernardini, a.barron}@unibo.it

Abstract

Spotting hate speech in social media posts is crucial to increase the civility of the Web and has been thoroughly explored in the NLP community. For the first time, we introduce a multilingual corpus for the analysis and identification of hate speech in the domain of inceldom, built from incel Web forums in English and Italian, including expert annotation at the post level for two kinds of hate speech: misogyny and racism. This resource paves the way for the development of mono- and cross-lingual models for (a) the identification of hateful (misogynous and racist) posts and (b) the forecasting of the amount of hateful responses that a post is likely to trigger. Our experiments aim at improving the performance of Transformer-based models using masked language modeling pre-training and dataset merging. The results show that these strategies boost the models' performance in all settings (binary classification, multi-label classification and forecasting), especially in the cross-lingual scenarios.

Disclaimer: Due to the nature of the topic, this paper contains offensive words.

1 Introduction

Hate speech can be generally defined as "language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" (Davidson et al., 2017). Detecting hate speech can be challenging as there is a lack of consensus on its definition, while the use of offensive neologisms makes the task even more arduous (Fortuna et al., 2020). This is even more critical in environments frequented by incels, short for *involuntary celibates*, which pertain to the so-called *manosphere* (Nagle, 2017, p. 75-86) and mainly comprise men unsuccessful in finding a sexual partner or significant other. Some of these individuals tend to engage in the spread of various forms of hate speech —in particular racism and misogyny— and recurrently adopt novel lexicon in doing so (Blommaert, 2018). Such dynamic jargon causes models trained on hate speech to fail to recognize incel-specific instances of hate speech.

Our contributions are the following:

(*i*) Corpora. We introduce two unsupervised corpora on the inceldom domain, one in English and one in Italian. A subset of each corpus includes manual annotations for different kinds of hate speech (cf. Section 3).[1] The raw data can be used for domain adaptation and language modeling, among other applications. The annotation allows addressing three tasks. *Binary:* determine whether a post p conveys hate speech or not. *Multi-label:* determine whether p is misogynous and/or racist. *Forecasting:* Given an original post p' (the first post in a thread), forecast the amount of hateful posts that it is likely to trigger in future responses.

(*ii*) Masked language modeling. We perform mono- and cross-lingual masked language modeling (MLM) to adapt BERT and mBERT models to the inceldom domain for the first time. We release the best configurations according to their impact in the identification of hate speech (Section 4).[2]

(*iii*) Hate speech identification. We show the impact of domain-adapted Transformers and the downstream training of models for hate speech identification, in the niche context of incel hate speech. We combine new incel-specific and existing supervised corpora within and across languages in three settings: binary classification, multi-label classification and forecasting (Section 5).

Our experiments show that MLM pre-training is effective, particularly in cross-lingual scenarios, resulting in a 17-point absolute improvement in

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

[2] The model configurations are available at: https://github.com/paolo-gajo/RANLP-2023-Models

Proceedings of Recent Advances in Natural Language Processing, pages 373–384
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_042

terms of F_1-measure in the binary task, and a 34- and 18-point increase in the misogyny and racism detection tasks, respectively. Combining Italian and English datasets leads to a large performance increase of 22 points in terms of F_1-measure, for the best MLM pre-trained model. In the forecasting setting, our regression model effectively predicts the number of hateful responses a post may generate in the following replies, surpassing the mean squared error (MSE) baseline by 37%.

2 Related Work

Corpora built from incel platforms are rare and not necessarily applicable to the use-case of this study, either due to the source of the data only being partially compatible with the linguistic domain presently tackled (Pelzer et al., 2021) or because of the criteria according to which it was annotated (Zhou et al., 2022). Most studies have focused on the linguistic properties of incel corpora, predominantly adopting qualitative approaches. For example, Tranchese and Sugiura (2021) compared incel discourse from Reddit forums to the language used in pornography and highlighted its misogynistic implications. Papadamou et al. (2020) conducted a cross-platform study on incel profiling, by collecting $6.5k$ YouTube videos shared by users in Incel forums within Reddit, while also examining the YouTube recommendation algorithm. Their findings show that incel activity on YouTube is increasing, stirring towards the dissemination of incel views. Jaki et al. (2019) adopted a mixed approach, mainly focusing on text profiling, with their discourse analysis suggesting that incel language is not as coherent as previously assumed, while also employing a multichannel CNN, using $50k$ Incels.me messages, $50k$ neutral texts composed of $40k$ paragraphs from random English Wikipedia articles, and $10k$ random English tweets. Past studies have relied on the Pushshift Reddit API to build a corpus within the linguistic domain of inceldom (Farrell et al., 2020; Mollas et al., 2022). Zampieri et al. (2019) build a dataset from English tweets which can be used to train models to identify and categorize offensive posts, with information on whether the target is a group or individual.

Recently, more hate speech studies turn towards a new approach: *forecasting*. Zhang et al. (2018) extract politeness strategies and rhetorical prompts to predict whether a conversation will turn uncivil. Meng et al. (2023) predict the intensity of hate that a tweet might carry through its reply chain by exploiting tweet threads and their semantic and propagating structures. Dahiya et al. (2021), compiled a dataset of $4.5k$ tweets and their reply threads, confirming that longitudinal patterns of hate intensity among reply threads are diverse, with no significant correlation with the source tweet. Their approach differs from ours in that they calculate hate intensity for chunks of a thread, not for the whole thread at once. Almerekhi et al. (2020) proposed a model for toxicity triggering prediction by integrating text-based features as well as features that are related to shifts in sentiment, topic flow, and discussion context, proving that toxicity triggers contain detectable features. Lin et al. (2021) proposed a model that uses a post's semantic, propagation structure, and temporal features to predict hateful propagation in social media.

3 Incel Corpora

We performed a *modern diachronic* study (Partington, 2010) on incel forums, shedding light on the way the language of inceldom evolves. We consider two forums: *Incels.is*,[3] in English, and *Il forum dei brutti*,[4] in Italian. Studying such niche communities, as opposed to those hosted for example on Reddit, allows us to study a language which is representative of the incel speech community. This is because moderation is more lax,[5] which allows users to express themselves more genuinely.

The study, discussed at length in Appendix A, shows that excessively outdated resources might not be entirely representative of the discourse currently produced by the speech communities being scrutinized. More worthy of notice is that incel language differs from general Internet language, especially when hate speech is expressed. Such findings show that building new corpora from scratch is a worthwhile effort, as having an accurate representation of current language is a priority.

We retrieved dumps of posts from the two forums. The metadata for each post includes: author id, the position of the post in the thread, URL, timestamp and both post and thread unique ids.

We refer to the unsupervised dataset obtained from the dump of the *Incels.is* forum as IFU-22-EN

[3] `https://incels.is` (Last access: 11 August 2023)

[4] `https://ilforumdeibrutti.forumfree.it` (Last access: 11 August 2023)

[5] The /r/incels and /r/braincels subreddits, the most popular to date, were respectively shutdown in 2017 and 2018 because of the hatefulness of their contents.

Dataset	Posts	Threads	Length
IFU-22-EN	4,7M	223k	31.07±70.01
IFU-22-IT	627k	30k	52.78±80.77

Table 1: Statistics of the IFU-22-EN and the IFU-22-IT unsupervised corpora (length computed in tokens).

Please identify whether each post is categorized as misogynous, racist, or falls into another category:
A post is deemed **misogynous** if it:

- 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; or
- Uses slurs against women purposelessly.

A post is considered **racist** if it:

- 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; or
- Shows support for problematic ideologies, such as xenophobia, homophobia, or sexism.

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

(Incel Forum Unsupervised, 2022, English). The posts it contains come from the "Inceldom Discussion" section. The dataset extracted from *Il forum dei brutti*, which we refer to as IFU-22-IT (Incel Forum Unsupervised, 2022, Italian), comes from the "Una vita da brutto" section. Table 1 shows the statistics of the two datasets. The average length of the posts is much longer in Italian than in English. The median posting time difference between an original post and its first response is also much higher in IFU-22-IT, with a median of 540 against only 155 seconds. This could hint that threads in *Il forum dei brutti* are less active as far as the frequency of replies is concerned, but hosting conversations which are more akin to actual discussions, rather than the more chaotic back-and-forths which seem to take place in *Incels.is*.

We annotated a subset of the posts from both collections with two independent binary labels: one for misogyny and one for racism. We refer to the resulting datasets as IFS-EN and IFS-IT, which stand for Incel Forum Supervised in English (with 5,203 instances) and Italian (with 500 instances).

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

Table 2: Class distribution for the IFS-EN and IFS-IT supervised datasets. Mis=misogynous, Rac=racist.

IFS-EN was initially sampled with two constraints: 50% of the posts had to include at least one term characteristic of incel jargon[6] and instances had to be longer than five words. The former constraint sought to balance the occurrence of instances with and without incel jargon to prevent models from overly relying on it, while the second aimed at excluding instances which would not be useful during training. For IFS-IT, only a 5-word minimum length constraint was applied. Figure 1 shows the annotation guidelines.

With relation to English, a pilot annotation was first carried out by three annotators on a subset of 50 instances. All annotators have a C2 CEFR level of English and are experts in the subject, with a strong foundation in linguistics and gender studies, as well as knowledge of NLP and data annotation. The obtained Cohen's Kappa inter-annotator agreement (IAA) (Bobicev and Sokolova, 2017) was of 0.77, considered *substantial* (with 0.81 being the threshold for *almost perfect*). The rest of the instances were annotated by a single annotator. As for Italian, two annotators, native speakers of Italian and with the same background as above, obtained an IAA of 0.69 over 50 instances. As the IAA was deemed acceptable, the 450 other instances were all labeled by a single annotator.

We split IFS-EN into training, development and testing partitions with a ratio of 70/15/15, while we use IFS-IT only for cross-lingual testing. Table 2 shows the statistics of the two supervised corpora. About 1.2% of the instances are judged as both misogynous and racist.

4 MLM Pre-Training

We build upon BERT base for monolingual English scenarios and mBERT base (Devlin et al., 2019) for cross-lingual scenarios in English and Italian. Based on Caselli et al. (2021), we attempt

[6] Used terms: shitskin, racepill, deathnic, stacie, cumskin, jb, noodlewhore, chadlite, slav, whitecel, foid, cunt, curryland, slut, aryan, deathnik, ricecel, roastie, whore, femoid. See Appendix A for details on the selection process.

	MLM	Validation (English)			Test (English)			Test (Italian)		
	Dataset	F_1	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec
Monoling.	BERT	0.846±0.010	0.851	0.845	0.845±0.008	0.843	0.849			
	EN 10k	0.867±0.005	0.870	**0.865**	0.865±0.008	0.855	**0.876**			
	EN 100k	0.865±0.006	0.887	0.846	0.868±0.006	0.882	0.855			
	EN 1M	**0.875±0.005**	**0.894**	0.856	**0.872±0.006**	**0.883**	0.861			
Cross-lingual	mBERT	0.843±0.005	0.862	0.826	0.826±0.007	0.803	0.851	0.333±0.114	0.224	0.742
	IT 10k	0.842±0.005	0.868	0.818	0.840±0.009	0.807	0.876	0.410±0.099	0.290	0.746
	IT 100k	0.847±0.005	0.862	0.834	0.836±0.007	0.809	0.865	0.249±0.089	0.150	0.804
	IT 627k	0.844±0.006	0.855	0.834	0.836±0.008	**0.819**	0.855	0.111±0.060	0.060	0.861
	EN 10k	0.854±0.006	0.882	0.827	0.837±0.005	0.797	0.881	0.501±0.050	**0.378**	0.762
	EN 100k	0.852±0.003	0.876	0.830	0.835±0.009	0.797	0.878	0.371±0.106	0.246	0.843
	EN 1M	0.859±0.006	0.882	0.837	0.835±0.005	0.789	0.888	0.112±0.034	0.060	0.857
	EN–IT 10k	0.847±0.009	0.863	0.833	0.831±0.004	0.806	0.858	0.179±0.060	0.102	0.831
	EN–IT 100k	0.852±0.007	0.882	0.825	0.824±0.007	0.783	0.871	0.341±0.079	0.221	0.793
	EN–IT 1M	**0.863±0.004**	**0.887**	**0.841**	**0.845±0.006**	0.801	**0.894**	**0.503±0.042**	0.356	**0.864**

Table 3: Impact of MLM training on the performance of mono- and cross-lingual hate speech binary classification.

Dataset		Source	Lan
Davidson	(Davidson et al., 2017)	Hatebase.org	en
HateXplain	(Mathew et al., 2021)	Twitter+Gab	en
Stormfront	(Mathew et al., 2019)	Stormfront.org	en
HatEval	(Basile et al., 2019)	Twitter	en
HSD_{fb}	(Bosco et al., 2018)	Facebook	it
HSD_{tw}	(Bosco et al., 2018)	Twitter	it

Table 4: Existing hate speech datasets used to enrich the binary classification models.

to improve the models' understanding of the incel language by training them on the MLM task, producing what we refer to as in-domain *Incel BERT* and *Incel mBERT* versions.

In the monolingual scenario, three samples from the IFU-22-EN unsupervised dataset are used, considering randomly-selected splits of $10k$, $100k$, and $1M$ posts. We adopt a similar approach in the cross-lingual scenario, where we consider (*i*) the same English subsamples alone; (*ii*) subsamples of $10k$, $100k$, and $627k$ instances in Italian from IFU-22-IT (the full corpus contains $627k$ instances); and (*iii*) 50–50% splits from both IFU-22-EN and IFU-22-IT of $10k$, $100k$, and $1M$ instances. None of the instances used for MLM pre-training include data from IFS-EN and IFS-IT.

In all cases, MLM pre-training is carried out by tokenizing posts with AutoTokenizer[7] and masking tokens with a probability of 15%. We use a batch size of 32 and train the models for one epoch on a single Tesla P100 GPU with 16 GB of VRAM.

In order to assess the impact of the MLM pre-training, we perform preliminary experiments on the binary classification task: hate speech or not. We fine-tune each model version using $\text{IFS-EN}_{\text{tr}}$ for training and $\text{IFS-EN}_{\text{de}}$ for development.[8] We then test on $\text{IFS-EN}_{\text{te}}$ in the monolingual scenario and on IFS-IT in the cross-lingual scenario. Our baseline for monolingual scenarios is BERT, while we use mBERT in cross-lingual ones.

Table 3 reports the results. The experiments are repeated ten times in order to make our results more reliable and diminish the effect of random initializations. As it is common (e.g., Pelicon et al. (2021); Muti and Barrón-Cedeño (2022)), mBERT achieves inferior results in the monolingual scenario compared to BERT. Pre-training BERT on $1M$ monolingual instances on the MLM task improves model performance and yields the best results. The performance improves linearly but subtly as more data is introduced, reaching a 3-point absolute difference: from 0.845 to 0.872. When zooming into posts which do not contain incel terminology, the performance of the models is lower, but pre-training on $1M$ monolingual instances still provides a performance boost over BERT_{base} (0.727 vs 0.671). When looking at posts which contain incel terminology, both models obtain an F_1 of 0.934, showing that explicit hate is much easier to detect.

In the cross-lingual scenario, MLM also has a positive impact, but the improvement is not linear with the amount of data. When performing MLM with monolingual data, be it in English or Italian,

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

[8] For all classification experiments, the number of epochs is set based on the performance achieved on $\text{IFS-EN}_{\text{de}}$.

Model	English				Italian		Validation (English)			Test (English)		
	Davidson	HateXplain	Stormfront	HatEval	HSD FB	HSD TW	F_1	Rec	Prec	F_1	Rec	Prec
BERT							0.846±0.010	0.851	0.845	0.845±0.008	0.843	0.849
	■						0.838±0.010	0.834	0.843	0.851±0.006	0.852	0.849
			■				0.853±0.008	0.854	0.852	0.855±0.005	0.863	0.848
			■	■			0.847±0.002	0.853	0.843	0.849±0.009	0.862	0.837
Incel BERT							**0.875±0.005**	**0.894**	0.856	**0.872±0.006**	0.883	0.861
		■					0.858±0.003	0.789	**0.940**	0.857±0.008	0.804	**0.918**
			■				0.859±0.004	0.861	0.858	0.865±0.004	**0.884**	0.848
			■	■			0.859±0.002	0.882	0.838	0.859±0.002	0.882	0.838
							Validation (English)			Test (Italian)		
mBERT							0.843±0.005	0.862	0.826	0.333±0.114	0.224	0.742
					■		0.835±0.010	0.837	0.835	0.694±0.011	0.859	0.583
						■	0.854±0.011	0.875	0.835	0.657±0.035	0.721	0.612
					■	■	0.825±0.005	0.780	0.876	0.690±0.012	0.807	0.605
Incel mBERT							0.863±0.004	**0.887**	0.841	0.503±0.042	0.356	**0.864**
					■		**0.862±0.002**	0.856	0.867	0.704±0.003	**0.893**	0.582
						■	0.859±0.007	0.886	0.834	0.695±0.023	0.641	0.764
					■	■	0.855±0.008	0.834	**0.877**	**0.721±0.010**	0.842	0.630

Table 5: Impact of incorporating additional datasets in English (Italian) when fine-tuning BERT (mBERT) and Incel BERT (Incel mBERT) on the mono- (top) and cross-lingual (bottom) hate speech detection task.

the testing performance on both languages is better than vanilla mBERT, when using $10k$ instances, but drops with additional monolingual training material. Using a bilingual combination of MLM material produces the best model when using $1M$ instances. This configuration boosts the performance: (*i*) by 39 points on Italian, with respect to adding $1M$ of all-English instances (0.503 vs 0.112) and (*ii*) by 1 point on the English one (0.845 vs 0.835). With respect to the mBERT baseline, training on $1M$ bilingual instances provides a performance boost of 17 points (0.503 vs. 0.333).

Going forward, we use the best post-MLM models: Incel BERT trained on $1M$ English instances in monolingual experiments and Incel mBERT trained on $1M$ bilingual instances in cross-lingual ones.

5 Downstream Tasks

This section discusses our three experimental settings: (*i*) binary hate speech classification, (*ii*) multi-label misogyny and racism classification, and (*iii*) hate speech forecasting. In all settings we tokenize input sentences with AutoTokenizer, padding to a maximum of 256 tokens, including [CLS] tokens, and returning attention masks. All models are trained with a batch size of 16, using the AdamW optimizer with $lr = 10^{-5}$ and $\epsilon = 10^{-8}$.

Both classification tasks are evaluated on the basis of F_1-measure. The forecasting (regression) task is evaluated using mean squared error (MSE) and mean absolute error (MAE).

5.1 Binary Hate Speech Classification

Following the approach of Pelicon et al. (2021), we enrich the models while training them on the downstream binary task by using various combinations of existing datasets labeled for hate speech, summarized in Table 4. The Davidson dataset is subsampled to the size of IFS-EN$_{\text{tr}}$ because doing so performed better in preliminary experiments. For HatEval, we only use the part pertaining to misogyny, as the instances annotated for hate speech against migrants were not relevant with relation to incel speech. Table 5 displays the results for the dataset combinations which performed the best.

Monolingual scenario. Combining IFS-EN$_{\text{tr}}$ with the Stormfront, Davidson, and Stormfront+HatEval datasets slightly improves BERT's performance, respectively yielding an improvement of 1, 0.6 and 0.4 points on the test set. Neither HatEval nor HateXplain contribute positively. In the case of HatEval, this is probably due to the fact that it focuses only on misogynous hate speech, which is not entirely representative of the problem at hand. As

	Label	Model	Validation (English)			Test (English)		
			F$_1$	Rec	Prec	F$_1$	Rec	Prec
Monoling.	M	BERT	0.759±0.009	0.737	0.783	0.804±0.014	0.800	0.808
		Incel BERT	0.786±0.005	0.786	0.786	0.803±0.005	0.826	0.782
	R	BERT	0.831±0.006	0.874	0.791	0.796±0.012	0.838	0.759
		Incel BERT	0.854±0.012	0.838	0.872	0.821±0.012	0.818	0.823
						Test (Italian)		
Cross-ling.	M	mBERT	0.764±0.022	0.749	0.781	0.214±0.102	0.127	0.813
		Incel mBERT	0.773±0.008	0.757	0.790	0.552±0.049	0.404	0.886
	R	mBERT	0.818±0.010	0.859	0.781	0.393±0.015	0.354	0.459
		Incel mBERT	0.828±0.007	0.876	0.786	0.577±0.045	0.523	0.644

Table 6: Results for the mono- and cross-lingual scenarios of the misogyny (M) and racism (R) classification setting.

regards HateXplain, it likely failed to improve the performance of the model because it was built to be used jointly with the attention arrays it contains and because its sentences are already tokenized and stripped of punctuation, which means the model has less syntactical information to work with.

As for Incel BERT, all combinations yielded worse results than the baseline. This could be because the model became too biased toward IFS-EN$_{\text{tr}}$, making it unable to learn effectively from other datasets. That said, Incel BERT's results on IFS-EN$_{\text{te}}$ are still better than the ones BERT achieves when merging IFS-EN$_{\text{tr}}$ with Stormfront, Davidson, or Stormfront+HatEval.

Cross-lingual scenario. As expected, despite the annotation schema of our datasets and the ones we add to them being different, providing mBERT with extra training material in Italian (HSD$_{\text{fb}}$ and HSD$_{\text{tw}}$) improves the model, compared to only fine-tuning on IFS-EN. All models improve over the baseline, reflecting the importance of adding training material in the target language, even if no MLM pre-training is carried out at all. The best performance is achieved when adding HSD$_{\text{fb}}$ alone, with a performance on par with that obtained when adding both datasets. The difference of 36.1 points hints at a high affinity between the annotation schemes of HSD$_{\text{fb}}$ and IFS-IT.

A similar trend can be observed when training Incel mBERT by also adding both HSD$_{\text{fb}}$ and HSD$_{\text{tw}}$ to the training data, with a 22-point increase (from 0.503 to 0.721). When evaluating on Italian, using both English and Italian for MLM training and merging both HSD$_{\text{fb}}$ and HSD$_{\text{tw}}$ to IFS-EN for fine-tuning outperforms the rest of the alternatives. This is the case even if departing from vanilla Incel mBERT, which performs the worst before adding Italian fine-tuning data.

In general, in both mono- and cross-lingual scenarios, a lower standard deviation is observed for Incel BERT and Incel mBERT when additional training material is added, reflecting that the models gain substantially in stability thanks to it.

5.2 Multi-Label Hate Speech Classification

In this case, we fine-tune for the multi-label problem of identifying misogynous and/or racist posts, again in mono- and cross-lingual scenarios.[9] In both cases, only IFS-EN is used for training and development. In the monolingual scenario, testing is done on IFS-EN$_{\text{te}}$, while in the cross-lingual scenario IFS-IT is used. Table 6 shows the results for each individual class.

Monolingual scenario. The misogyny detection performance obtained by BERT and Incel BERT 1M on the Italian test set is essentially the same: 0.803 vs 0.804 F$_1$-measure. Incel BERT's recall is better than vanilla BERT's, which could reflect that MLM pre-training is indeed helping the model identify misogyny more effectively, but at the same time turning it more permissive.

Regarding racism, Incel BERT performs slightly better than BERT, with an absolute difference of 2.5 points: 0.821 vs 0.796. Just like in the binary setting, the performance boost obtained by Incel BERT is the result of the model already being familiar with the novel racist language used by incels.

Cross-lingual scenario. Both with relation to misogyny and racism identification, the performance of Incel mBERT is far higher compared to vanilla mBERT's. As far as misogyny identification is concerned, Incel mBERT outperforms the baseline mBERT model by 33.8 points, while in the racism detection task it outperforms the baseline by 18.4 points. These results suggest that using target

[9] In this setting, each model is fine-tuned five times.

Corpus	HS	(%)	No HS
IFU-22-EN	836,974	(17.59)	3,919,908
IFU-22-IT	282,724	(44.30)	355,419

Table 7: Class distribution of the predicted labels on IFU-22-EN and IFU-22-IT, showing the number of posts judged as being hate speech (HS) or not (No HS).

language data for MLM pre-training can greatly increase the performance of a model even without using any target language (Italian, in this case) data for fine-tuning on the downstream task.

As opposed to the monolingual scenario, in this case the greater performance boost is also an indication that exposing the model to the target language domain is highly effective. This shows that the language of inceldom in *Il forum dei brutti* is indeed very different from general Italian language, in line with the diachronic study of Appendix A, and that the model benefits from learning its features.

5.3 Hate Speech Forecasting

In the context of an Internet forum, we define forecasting as the capability of predicting how many posts will contain hateful content following an original post p' as soon as it has been posted. We conceptualize the amount of hate generated in a thread as the ratio between the number of hateful posts following p' and the total number of posts contained in the thread it has started. Based on this rationale, we build two corpora, one in Italian and the other in English, in which each p' is paired to a *hate score* in the range $[0, 100]$, indicating how much hate it has generated, with the extremes representing that none or all of the thread's posts are considered hateful.

To produce the data for this setting, we first generate automatic binary predictions for all the posts in IFU-22-EN and IFU-22-IT using the top models from Section 5.1: Incel BERT trained on IFS-EN alone for the former and Incel mBERT trained on IFS-EN$_{\text{tr}}$ plus HSD$_{\text{fb}}$ and HSD$_{\text{tw}}$ for the latter. Table 7 shows the resulting class distribution, which is in line with the training material's. We use these binary decisions to compute a silver hate score for each p' in the corpora. The resulting collection of p'–hate score pairs in English includes $223k$ instances, while the Italian one has $30k$.

Figure 2 shows histograms of the hate score distributions in both languages. The distribution for English is skewed to the left, with a median of 13.89, indicating that most original posts tend to trigger a small amount of hateful responses. The Italian distribution resembles a Gaussian with a median of 42.86, except for the outliers at the extremes. This reflects a uniform range in the amount of hate triggered by comments in the Italian forum.

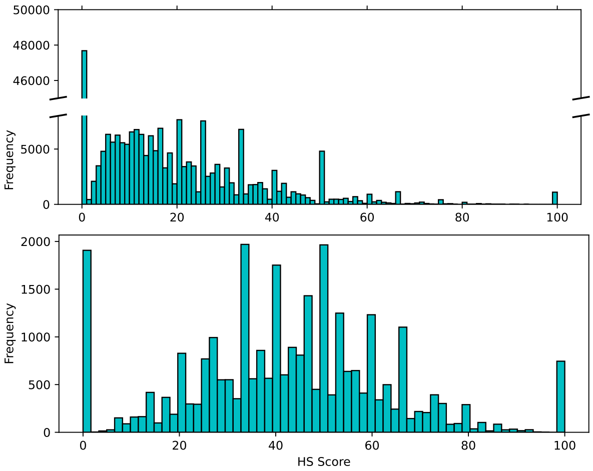

Figure 2: Hate score distribution associated to the original posts for English (top) and Italian (bott.) forecasting.

It is clear that many of the original posts in both forums trigger no hate, while a smaller number triggers a plethora of hateful responses. The number of completely non-hateful threads is much higher in the English OP–score corpus while, comparatively, the number is much lower in the Italian one, where it is on par with the center of the distribution. As regards the number of threads with a hate score of 100, the opposite is true: *Il forum dei brutti* has a much higher percentage of hate, because in most of its threads which only have one reply, that reply is hateful (515 out of 921 single-reply threads).

We address forecasting as a regression problem and train Incel BERT and Incel mBERT to output continuous $[0, 100]$ hate scores. We do this by adding a 1D linear output layer on top of them. Unlike previous experiments, here we train the models only on original posts p' and for a different objective. We split both English and Italian corpora into training, development and test sets with ratios of 70/15/15 and use them to train and evaluate mono- and cross-lingual models. Following the approach of Kang et al. (2018), our baselines are the means of the scores contained in the development and test partitions of the produced hate score datasets.

Table 8 shows the results, recorded over four epochs. We set the maximum number of epochs at four because in the cross-lingual scenario the tuning converges on the fourth epoch.

Monolingual scenario. The model performs better

e	**Monolingual**			**Cross-lingual**		
	MSE$_{va}$	**MSE$_{te}$**	**MAE**	**MSE$_{va}$**	**MSE$_{te}$**	**MAE**
1	**188.63**	**181.19**	9.95	590.98	586.65	19.37
2	192.71	186.28	10.36	466.27	462.58	16.71
3	195.50	188.51	9.94	436.57	432.68	16.12
4	203.52	196.25	10.24	**425.13**	**421.70**	15.95
b	296.18	286.44	13.17	461.84	457.47	16.56

Table 8: Performance in terms of MSE (val. and test) and MAE (test) for the forecasting task, for the mono- and cross-lingual scenarios; e=epoch, b=baseline.

than the baseline right from the first epoch, on which it achieves its top performance with an MSE of 181.19, 36.74% lower than the baseline. This indicates that the model is reasonably effective at forecasting the amount of hate that an original post is going to generate. This is also supported by the fact that, for instance, the mean absolute error (MAE) on the English test set after one epoch is 9.95, compared to 13.17 for the baseline.

Cross-lingual scenario. Incel mBERT struggles more at forecasting, with the best MSE on the Italian test set being 421.70, which corresponds to a MAE of 15.95. Compared to the monolingual scenario, the performance gap from the baseline is also not as significant (7.82%). Other than the difficulty added by the cross-lingual component, the noisier silver data produced by a lower-performing single-post classification model makes effective forecasting more challenging, which is also reflected by the slow convergence after additional epochs.

These results, particularly those in the monolingual setting, hint that it would be possible to estimate the amount of hate that a post is likely to trigger —just by looking at its textual content— as soon as it has been posted, although the prediction quality has room for improvement.

6 Conclusions

In this paper, we have explored the creation of models for the automatic identification of hate speech in incel forums: binary hate speech identification, multi-label misogyny and racism identification, and forecasting of the level of hate that the first post of a thread is likely to trigger.

Our experimentation on the three problems, in monolingual and cross-lingual scenarios, shows that (*i*) pre-training on the masked language modeling task to make BERT-based models more aware of incel language is a key factor to aspire to produce good predictions; (*ii*) the inclusion of supervised material extracted from sources external to incel forums can help boost models further, also across languages; and (*iii*) it is feasible to forecast the amount of hate that an original post will likely trigger prior to any replies, although further improvements are still required.

In future work, we plan to delve further into forecasting by implementing temporal and propagation features (e.g., Meng et al. (2023); Dahiya et al. (2021); Lin et al. (2021); Almerekhi et al. (2020); Jaki et al. (2019)). Based on Pelicon et al. (2021), we also plan to expand language coverage, with German- (Mandl et al., 2019) and Spanish-language (Basile et al., 2019) hate speech datasets being two of the most prominent candidates due to their similarity to English and Italian, respectively.

Limitations

The large amount of explicit hate in the training data might lead the models to prioritize detecting overtly offensive language while potentially overlooking more subtle forms of implicit hate. Consequently, instances of implicit hate within threads might be misclassified, affecting both the classification and regression-based evaluation.

We attempted to assess our models' generalizability with preliminary cross-domain tests on the Contextual Abuse Dataset (Vidgen et al., 2021). This was the only relevant thread dataset available, but its abusive language labels did not align with ours. The limited availability of thread datasets hindered further cross-domain and cross-lingual experiments, rendering further research timely.

The forecasting setting, built on top of post-level silver data as a proxy, could benefit from human annotation at the thread level. Still, making this task practical at scale is complex and expensive.

Ethical Considerations

All data used to compile our corpora is publicly available. Forum users accept a legal disclaimer before posting and are kept anonymous.

The paper covers sensitive topics which could be subject to bias and human supervision is necessary to assess the quality of the results, especially during the annotation process. Therefore, the annotated posts were evaluated as objectively as possible.

Although we reckon freedom of speech as a fundamental right, we advocate for online content moderation, given the real-world violence triggered by hate speech, as discussed in the introduction.

References

Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2020. Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference 2020*, WWW '20, page 3033–3040, New York, NY, USA. Association for Computing Machinery.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Jan Blommaert. 2018. Online-offline modes of identity and community: Elliot Rodger's twisted world of masculine victimhood. In *Cultural practices of victimhood*, pages 193–213. Routledge, Abingdon, Oxfordshire, UK.

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

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. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Snehil Dahiya, Shalini Sharma, Dhruv Sahnan, Vasu Goel, Emilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. Would your tweet invoke hate on the fly? forecasting hate intensity of reply threads on twitter. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2732–2742, New York, NY, USA. Association for Computing Machinery.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tracie Farrell, Oscar Araque, Miriam Fernandez, and Harith Alani. 2020. On the use of jargon and word embeddings to explore subculture within the reddit's manosphere. In *12th ACM Conference on Web Science*, WebSci '20, page 221–230, New York, NY, USA. Association for Computing Machinery.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

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

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.

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

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proceedings of Corpus Linguistics Conference CL2009*, University of Liverpool, UK.

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

Ken-Yu Lin, Roy Ka-Wei Lee, Wei Gao, and Wen-Chih Peng. 2021. Early prediction of hate speech propagation. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 967–974.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. 2023. Predicting hate intensity of twitter conversation threads. *arXiv preprint arXiv:2206.08406*. Accepted in Knowledge-Based Systems.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16.

Arianna Muti and Alberto Barrón-Cedeño. 2022. A checkpoint on multilingual misogyny identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 454–460, Dublin, Ireland. Association for Computational Linguistics.

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

Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2020. Understanding the incel community on youtube. *CoRR*, abs/2001.08293.

Alan Partington. 2010. *Modern Diachronic Corpus-Assisted Discourse Studies: Corpora Volume 5, Number 2*. Edinburgh University Press.

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.

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

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*.

Alessia Tranchese and Lisa Sugiura. 2021. "i don't hate all women, just those stuck-up bitches": How incels and mainstream pornography speak the same extreme language of misogyny. *Violence Against Women*, 27(14):2709–2734. PMID: 33750244.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing cad: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303.

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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. 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.

Appendix

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 terms, showing how the lexical features of incel 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 far as *Incels.is* is concerned, 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. Racism and misogyny are very characteristic elements of the language of incels (Silva et al., 2016; Ging and Siapera, 2018; Jaki et al., 2019). Therefore, we manually selected characteristic hateful terminology for this speech community by considering racist and misogynous terms that are not typically found in general language, i.e. having high keyness scores.

In order to conduct the diachronic study, the subset was divided into 22 chronological partitions, one for each 100 pages[10] of the forum from 2017 to 2022. The keyness of each selected term was measured for every partition, calculating the slope of its regression line across all 22 partitions. For each term, the slope was divided by the average keyness over the 22 partitions, thus obtaining its normalized slope. For each partition, only the terms having the top 500 keyness scores were recorded. Zero values (7.16% in total), 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 for the average keyness. The 10 terms with the highest and lowest normalized slope, 20 in total, were thus grouped, calculating their mean normalized slope.

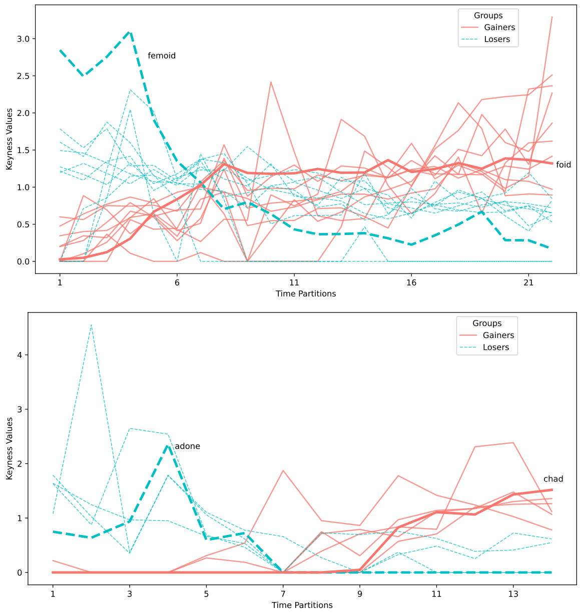

Figure 3: Keyness over time for the characteristic incel terms extracted from *Incels.is* (top) and *Il forum dei brutti* (bottom). Red (blue) lines represent the terms that gained (lost) keyness over time.

As regards *Il forum dei brutti*, the forum contents were divided chronologically by grouping posts by year of creation, from 2009 to 2022, for a total of 14 partitions. In this case, we carry out a study on 10 terms we deem to be characteristic of the forum's incel language, used to describe other men in negative or positive ways. The amount of zero values for these 10 terms is 44.44% of the total.

Figure 3 shows the over-time trend of the keyness of the terms extracted from *Il forum dei brutti* and *Incels.is* 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, "adone" is a close synonym of "chad", while "foid" is a contraction of "femoid", and for both pairs we can observe opposite trends with a specific point in time in which one overtakes the other.

Table 9 reports the normalized slopes of the

[10] Each page contains 10 threads.

	Gainer	**Slope**	**Loser**	**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**
FdB	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 9: Keyness normalized slopes for *Incels.is* and *Il forum dei brutti* (FdB).

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 *Il forum dei brutti*, the difference is 0.247, while for *Incels.is* the difference between the mean normalized slopes is smaller, 0.106, which points at a slower lexical evolution. For both forums, 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.

As regards *Il forum dei brutti*, we can observe that the way users refer to men changes in a rather clear way. Positive words that are commonly used in general language, such as "strafigo" (meaning "extremely handsome"), are substituted by specialized terms that are more specific to the forum's speech community, e.g., "chad". [11] Conversely, we can see the same phenomenon for negative words, where "reietto" ("outcast") loses popularity, leaving space to terms with more specialized uses. An example of this is "bv", meaning "brutto vero" (lit. "truly ugly"), which, being an acronym, is more opaque to outsiders.

With relation to *Incels.is*, as already anticipated through Figure 3, although terms like "foid" and "femoid" have the same meaning (both are used to dehumanize women by associating them to insentient androids), [12] the shorter form has become more popular, while the use of the full form has decreased. 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.

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. Therefore, if the material used to train models is outdated, their understanding of the discourse currently produced by a specific community could become suboptimal.

In both scenarios, it is thus arguably 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.

[11] https://incels.wiki/w/Chad (Last access: 11 August 2023)

[12] https://incels.wiki/w/Femoid (Last access: 11 August 2023)