# On the Identification and Forecasting of Hate Speech in Inceldom

**Anonymous ACL submission**

## 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 multilingual models for (a) the identification of hateful posts (binary and multi-label setting) and (b) the forecasting of the amount of hateful responses that a post is likely to trigger (regression setting). Our models reach an $F_1$ score above 0.85 in the classification settings and MAEs around 0.10 for the forecasting settings. These performances show that it is doable to approximate the extent of hate speech that a full thread is likely to contain, as soon as the first post has been made public —be it In English or Italian.

**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 in recognising Incel-specific instances of hate speech.

We have produced a multilingual —English and Italian— corpus on the inceldom domain that allows to address three tasks:

**Binary.** Given a post $p$, determine whether $p$ conveys hate speech or not.

**Multi-label.** Given a post $p$, determine whether $p$ is misogynous, racist, both, or neither.

**Forecasting.** Given a main post $p'$, forecast the amount of hateful posts that it is likely to trigger in future responses.

We explore with hate-tuned transformers and CNNs with incel-specific embeddings. [1] Our classification models reach $F_1$ performance values above 0.85 whereas our forecasting models reach MAE values around 0.10.

For the cross-language setting, we opt for a zero-shot approach from English to Italian to assess the capabilities of multilingual BERT in this domain and task. The outcome is ambivalent: showing a competitive performance in the binary setting, but significantly dropping in the multi-label one.

## 2 Related Work

Datasets 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, mostly adopting qualitative approaches. For example, Tranchese and Sugiura (2021) compared incel

[1] The corpus and the implementation are available at `https://blind.for.review.com` (submitted as supplementary material).

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 Convolutional Neural Network, using 50,000 Incels.me messages, 50,000 neutral texts composed of 40,000 paragraphs from random English Wikipedia articles, and 10,000 random English tweets. achieving an statistical accuracy of 95%. 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). Recently, more hate speech studies turn towards a new approach, that is *forecasting*. Meng et al. (2022) 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. Almerekhi et al. (2020) proposed a neural network for toxicity triggering prediction by integrating text-based features as well as features that related to shifts in sentiment, topic flow, and discussion context, proving that toxicity triggers contain detectable features. Lin et al. (2021) produced a deep learning model that uses a post's semantic, propagation structure, and temporal features to predict hateful propagation in social media which manages to outperform the best baselines by more than 10% (F1 and accuracy score).

## 3 Corpora and Tasks Definition

The rationale for a new English-language corpus within the inceldom sphere is based on a diachronic study of keywords characteristic of incel language over a 104M word subset of the messages posted on the "Inceldom Discussion" section of https://incels.is up to 18 October, 2022. The study sheds light on the way the lexicon of this community evolves. We produced 22 chronological partitions from 2017 to 2022 and measured the keyness (Kilgarriff, 2009) of terms among the partitions. Table 1 shows the normalised slopes of keyness for the top-10 gainers and bottom-10 losers among the characteristic incel lexical items over the 22 partitions (ignoring 0 values, 7.16% in total). The mean normalised slope of the top-10 gainers is 0.063, while it is -0.043 for the bottom 10 losers. This shows a clear upward trend for the gainers and downward trend for the losers, indicating a shift of lexicon usage within the lifetime of the forum. A change over time of the keyness of incel jargon is thus indication that the lexicon of dated resources is not fully representative of the current discourse involving inceldom, which means a new corpus was deemed necessary.

Table 1: Normalised slopes of the keyness of the top-10 gainers and bottom-10 losers among the lexical items characteristic of incel language.

| gainers | slope | losers | slope |
|---|---|---|---|
| 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** |

For the English partition, we crawled all messages posted on the "Inceldom Discussion" section of https://incels.is up to 18 October, 2022. We obtained a dump of $4.76M$ posts organised in $230k$ threads. For the Italian partition, referred to as IFC-22-it , we crawled the "*Una vita da Brutto*" section of https://ilforumdeibrutti.forumfree.it up to 4 December, 2022. We obtained a dump of 638k posts organised in 30k threads. IFC-22-it serves to observe whether multilingual transformers generalise well across languages when predicting incel-generated hate speech. For both languages, a post contains the text of the author and explicit quotations to previous posts in the thread are stored separately. The metadata includes author id, the position of the post in the thread, title, URL, timestamp and both post and

Table 2: Statistics of the IFC-22 corpus of hate speech in incels posts. The Italian partition, used only for testing, appears in the last row.

| **partition** | **misogyny** | **racism** | **none** |
|---|---|---|---|
| training | 797 | 620 | 2,179 |
| dev | 171 | 133 | 467 |
| test | 171 | 132 | 467 |
| test$_{\text{it}}$ | 98 | 6 | 149 |

Table 3: Statistics of the predicted labels on the IFC-22-en subset (100k instances) and on IFC-22-it .

| **corpus** | **misogyny** | **racism** | **none** |
|---|---|---|---|
| IFC-22-en | 9,393 | 4,049 | 76,074 |
| IFC-22-it | 8,119 | 4,288 | 614,001 |

thread unique ids.

We randomly sampled a subset of $5k$ posts in English for expert annotation considering three classes: misogyny, racism, or none. Two constrains were applied to select the posts: (a) length between 140 and 280 characters (so as to resemble the length of tweets) and (b) 50% of the posts had to include at least one characteristic word from the incel jargon in Table 1. The latter intends to guarantee a proportionate amount of instances with and without characteristic incel jargon, so as to prevent models from relying too much on them.

Three annotators expert in hate speech were recruited to perform the annotation. They were asked to follow the guidelines in Figure 1 to decide whether each post was misogynous, racist, a combination of both or neither of them. During a pilot annotation, all three volunteers annotated 50 instances independently (without considering any thread context). The kappa inter-annotator agreement (Bobicev and Sokolova, 2017, p. 100) among the three annotators was of 0.77, which corresponds to a "substantial" agreement, nearing "almost perfect". The rest of the instances were labeled only by one annotator.

For the Italian partition we manually annotated 250 instances following the same procedure as for English, resulting in 101 hate speech and 149 non-hate speech instances, with 98 being misogynous and only 6 being racist.

These manual annotations represent the gold standard for the binary and multi-label settings. Table 2 shows the statistics of both the English and the Italian partitions, which we refer to as IFC-22-en and IFC-22-it .

As for the forecasting task, —predicting the amount of hate that a main post is likely to trigger—, we used both the CNN and the BERT model to perform binary decisions over all the posts within the first 5,950 threads of the English forum in chronological order (89k posts in total).[2] We estimate the amount of hate that a main post produces as the ratio between the number of posts identified as hateful by the model and the total number of posts in its thread. The main posts and their ratios (i.e., the scores) represent the forecasting dataset. We then split the $6k$ instances into train (4165), val (892), and test (893) partitions.

## 4 Models Description

We performed experiments with both a CNN and a transformer architecture. In order to produce the representations for the CNN, we applied NLTK's casual_tokenizer[3] and built a 100D word2vec embedding space (Mikolov et al., 2013) using the $4.7M$ posts extracted from the incels.is forum. The CNNs are built with Keras[4] using a CNN layer with 16 filters and a kernel size of 3, global max-pooling, and a fully-connected layer with 250 neurons and ReLU activation function. We used a sigmoid function for classification. We train them using a batch size of 16 during 3 epochs using the Adam optimizer [5] over a binary cross entropy loss function and a dropout of 0.3. The architecture was identical in the multi-label setting, since we approached it as a binary relevance problem (Zhang et al., 2017). In order to turn the CNN competitive, we applied an active-learning approach (Hino, 2020), iteratively adding the top-10 and bottom-10 instances according to the model scoring up to reaching KKK instances. In the forecasting setting, the CNN uses mean absolute error as loss function.

For the transformer architecture, we use *bert-base-uncased-hatexplain* (Mathew et al., 2021), a version of BERT trained on Twitter and Gab hateful posts for the mono- and *bert-base-multilingual-cased* for the cross-lingual settings. In the

[2] In the monolingual (cross-language) setting, we used the top-performing BERT (mBERT) models for the binary tasks. The labels adopted were the ones that trained the top-performing models.

[3] https://www.nltk.org/_modules/nltk/tokenize/casual.html

[4] https://keras.io

[5] https://keras.io/api/optimizers/adam/

Table 4: Performance of the CNN and BERT models on the multi-label (left), binary (centre) hate speech identification tasks at the post level, and the forecasting (right) settings. Results on development and testing partitions included.

| | misogyny | | | racism | | | binary | | | forecasting |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **P** | **R** | **F$_1$** | **P** | **R** | **F$_1$** | **P** | **R** | **F$_1$** | **MAE** |
| CNN$_{\text{de}}$ | 0.87 | 0.87 | 0.87 | 0.86 | 0.85 | 0.85 | 0.87 | 0.86 | 0.87 | 0.14 |
| CNN$_{\text{te}}$ | 0.84 | 0.83 | 0.83 | 0.81 | 0.83 | 0.82 | 0.87 | 0.86 | 0.86 | 0.14 |
| BERT$_{\text{de}}$ | 0.80 | 0.80 | 0.80 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.13 |
| BERT$_{\text{te}}$ | 0.86 | 0.85 | 0.85 | 0.87 | 0.87 | 0.87 | 0.90 | 0.89 | 0.89 | 0.13 |
| mBERT(en) | 0.78 | 0.76 | 0.77 | 0.74 | 0.80 | 0.77 | 0.85 | 0.84 | 0.84 | 0.09 |
| mBERT(it) | 0.73 | 0.56 | 0.51 | 0.63 | 0.73 | 0.67 | 0.71 | 0.64 | 0.63 | 0.10 |

cross-lingual setting, the bert-base-multilingual-cased BERT model was fine-tuned on the IFC-22-en dataset containing 5,950 main posts. No preprocessing is applied to the text, other than applying the BertTokenizer. We use the AdamW optimizer with eps=1-8 and greedily search for the optimal epoch number with a held-out strategy in range $[1, 4]$ and a batch size of 16. For the binary task, we use the sigmoid activation function for the output layer. For the multi-label task we adopt a binary relevance approach (Zhang et al., 2017), combining two binary classification models. The output for each classifier is a sigmoid function too. We adopt this approach following Muti et al. (2022), since they show that treating the classes separately increased the performance when predicting misogynous, misogynous-aggressive or none. This approach allows us to predict mutually non-exclusive classes. For the forecasting task, we implement a regression model with BERT. The architecture is identical, but using regression layer with a BCE loss function at the end.

## 5 Experiments and Evaluation

We perform experiments aligned with the tasks introduced in Section 1 and considering the data partitions in Table 2.

Table 4 shows the results for both the mono- and the cross-lingual multi-label misogyny–racism–none and the binary hate speech or not settings, in terms of macro-avg F$_1$ score. The best performance on the test set is consistently obtained with the transformer models, increasing by 3 points over the CNN in the binary tasks, and by 2 and 5 points in misogyny and racism detection respectively.

In the zero-shot cross-lingual setting (last row), we observe a drop in the performance in both the binary and multi-label settings, which are likely due to the language-specific jargon used by incels. The drop might then suggest that the way incels express misogyny and racism is different across languages. Further studies are necessary to confirm this fact.

The last column of Table 4 shows the results of the forecasting setting. In the monolingual task, both the CNN and the BERT model obtained low MAE results, with BERT performing slightly better. In the cross-lingual setting, the model performs better on the English dataset, but almost as well on the Italian one. Actually, the performance is better than for the monolingual setting. This would be due to the fact that a larger share of posts in the Italian corpus are considered as non-hate speech (cf. Table 3), making the problem simpler.

## 6 Conclusions

We presented a novel corpus annotated for hate speech (misogyny and racism) derived from two inceldom forums: `incels.is` for English and `ilforumdeibrutti.forumfree.it/` for Italian. The corpus opens the door for the development of mono- and multilingual models for binary and multi-label prediction tasks, as well as forecasting regression tasks. Our experiments show that a transformer architectre outperforms a CNN baseline in all three tasks.

In future work, we would like to delve further into forecasting by implementing more large models and by comparing our results with state-of-the-art (Meng et al., 2022; Dahiya et al., 2021; Lin et al., 2021; Almerekhi et al., 2020; Jaki et al., 2019).

## Limitations

We tested the generalisability of our models by implementing preliminary cross-domain experiments on the Contextual Abuse Dataset (Vidgen et al.,

2021) in a binary setting, where the model obtained a 0.26 F1 score. Additionally, we evaluated the generalisability across languages, achieving relatively low results. The sparsity of datasets containing threads prevented us from performing more cross-domain ad cross-lingual experiments rendering further research timely. The only available thread dataset (Vidgen et al., 2021), that was relatively close to our purposes, contains gold labels for types of abusive language, and which did not match our dataset labels exactly. Therefore, we would like to encourage the creation of more such datasets that would allow a better model evaluation.

## Ethical Considerations

All the data for the compilation of the corpus were publicly available after going through a legal disclaimer. The posts are kept anonymous, as well as the posters maintain complete ownership of their posts.

The scope of the paper covers an inherently sensitive issue that could be subject to bias. Yet, we believe that human moderation is necessary to assess the quality of the results and especially during the annotation process, therefore, the annotated posts were evaluated with as much objectivity as possible.

## 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.

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.

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. arXiv.

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. Ibereval@ sepln, 2150:214–228.

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.

Hideitsu Hino. 2020. Active learning: Problem settings and recent developments. CoRR, abs/2012.04225.

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.

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

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.

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.

Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. 2022. Predicting hate intensity of twitter conversation threads.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

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, Francesco Fernicola, and Alberto Barrón-Cedeño. 2022. Misogyny and aggressiveness tend to come together and together we address them. In

Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4142–4148, Marseille, France. European Language Resources Association.

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.

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

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, Online. Association for Computational Linguistics.

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.

Min-Ling Zhang, Yukun Li, Xu-Ying Liu, and Xin Geng. 2017. Binary relevance for multi-label learning: an overview. Frontiers of Computer Science, 12:191–202.

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 Appendix

Specify whether each of the following posts is misogynist, racist, or otherwise.
A post is considered **misogynistic** if it:

- stereotypes or objectifies women;
- asserts superiority of men over women;
- derails the conversation to justify the abuse of women, reject male responsibility, disrupt the conversation to redirect it in favor of men;
- entails sexual advances, requests sexual favors, sexually harasses the recipient, aims to physically assert power over women through threats of violence;
- slurs women with no other purpose.

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

- uses a racial slur;
- negatively stereotypes, attacks, seeks to silence or criticises a minority without a founded argument;
- promotes violent crime against minorities;
- misrepresents the truth or distorts views on a minority with unfounded claims;
- shows support of problematic ideologies, e.g., xenophobia, homophobia, sexism.

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