Human-in-the-Loop Hate Speech Classification in a Multilingual Context

Ana Kotarcic[1,2], **Dominik Hangartner**[2], **Fabrizio Gilardi**[1],
Selina Kurer[2] and **Karsten Donnay**[1]

[1]Department of Political Science, University of Zurich, Switzerland
[2]Immigration Policy Lab, ETH Zurich, Switzerland
ana.kotarcic@gess.ethz.ch, dominik.hangartner@gess.ethz.ch, gilardi@ipz.uzh.ch,
selina.kurer@gess.ethz.ch, donnay@ipz.uzh.ch

Abstract

The shift of public debate to the digital sphere has been accompanied by a rise in online hate speech. While many promising approaches for hate speech classification have been proposed, studies often focus only on a single language, usually English, and do not address three key concerns: post-deployment performance, classifier maintenance and infrastructural limitations. In this paper, we introduce a new human-in-the-loop BERT-based hate speech classification pipeline and trace its development from initial data collection and annotation all the way to post-deployment. Our classifier, trained using data from our original corpus of over 422k examples, is specifically developed for the inherently multilingual setting of Switzerland and outperforms with its F1 score of 80.5 the currently best-performing BERT-based multilingual classifier by 5.8 F1 points in German and 3.6 F1 points in French. Our systematic evaluations over a 12-month period further highlight the vital importance of continuous, human-in-the-loop classifier maintenance to ensure robust hate speech classification post-deployment.

1 Introduction

Hate speech, often taken to designate insults and attacks against individuals or groups based on their inherent traits, is an expression which is widely used but on whose definition there is no general consensus. Given the subjective and contextual nature of hate speech, it is left open to media platforms (Meta, 2022; Twitter, 2022; Youtube, 2022; Microsoft, 2022b), research groups (Djuric et al., 2015; Saleem et al., 2017; Mondal et al., 2017; Salminen et al., 2018; Jaki and De Smedt, 2019; Pereira-Kohatsu et al., 2019; Rani et al., 2020; Röttger et al., 2021) and individuals to decide what to include under this notion and whether to adjust its definition (Vengattil and Culliford, 2022). Particularly difficult is drawing the line between hate speech, toxic speech and humour expressed through e.g. irony, sarcasm or euphemisms.

The challenges described in this paper show that configuring a stable and robust hate speech classifier is not a trivial task. While this applies to big tech companies who have the data and infrastructure to create, train and maintain their own classification systems, it is often prohibitive for smaller, mostly regional, companies and (online) media outlets. For them, no tailored off-the-shelf solution exists, they often do not have the in-house capacity to develop and maintain their own classification system and are bound by data protection rules tightly regulating which data can be shared with commercial services. The Swiss context presents a particularly challenging case given that many instances of hate speech are deeply ingrained in the multilingual, sociocultural and political contexts of Switzerland (cf. e.g. Stevenson, 1990; Hega, 2001; Brügger et al., 2009; Mueller and Dardanelli, 2014; FDFA, 2020) and are not picked up by currently available toxic speech classifiers. This situation is exacerbated by the fact that, up to the point of completing this study, no annotated hate speech data was available on which a solid baseline classifier for the Swiss context could have been trained.

The classification framework we developed addresses an important research gap regarding long-term, stable hate speech classification and is directly motivated by how such a classifier would be deployed in practice. A core feature of this system is that the human analysts remain both in control and at the centre of the decision making process, which we believe is vital to ensuring good post-deployment performance and enabling classifier maintenance. In building our human-in-the-loop hate speech classification pipeline, we closely collaborated with several Swiss media outlets and a Swiss NGO working on countering online hate.

In what follows, we will describe in more detail the steps involved in constructing this pipeline from the data collection stage all the way to post-deployment testing and the challenges encountered

in the process. Due to non-disclosure agreements signed with our partners, we are not able to disseminate our dataset or code, but we can share the technical details of our framework and approach. We also include an extensive appendix describing the nature of our dataset and details relating to the annotation process.

2 Related Work

Research on online hate speech and its detection has risen exponentially over the past few years, yielding a wide variety of approaches to tackling the phenomenon. Given the large body of work, numerous surveys and overviews outlining the main directions of research have been published on the subject (Schmidt and Wiegand, 2017; Salminen et al., 2018; Fortuna and Nunes, 2018; Vidgen et al., 2019; Abro et al., 2020; MacAvaney et al., 2019; Siegel, 2020; Mohiyaddeen and Siddiqi, 2021; Wenjie and Arkaitz, 2021; Mullah and Zainon, 2021; Tontodimamma et al., 2021). The research they review has mostly focused on optimising classifier performance by focusing on different types of input data, preprocessing steps, languages, classifier types, model architectures and feature engineering options, and by performing classification on various types of problematic language and target groups (cf. Table 8 in Appendix B).

Additionally, a series of shared tasks on offensive, abusive and hate speech classification have featured in NLP and linguistics competitions (cf. Table 9 in Appendix B) and are usually carried out in controlled environments with carefully configured datasets. Tasks are typically divided into two main categories: binary classification and multiclass classification, both for specific target groups and different types of offensive language.

The results of these challenges highlight that both the model selection and the nature of the dataset have a direct influence on classification performance whereby transformer models (Vaswani et al., 2017), particularly those based on the BERT architecture (Devlin et al., 2019), significantly outperform other models. Multiclass classification approaches perform consistently worse than do binary ones. When more than one language is present in the dataset, classification is usually not carried out in a multilingual setting but for each language individually. Classifiers for English often perform much better due to the sheer number of data points and pretrained models available, reaching F1 scores

of 83.0 for binary and 66.6 for multiclass classification (Mandl et al., 2021). In contrast, the best-performing German language classifiers reach F1 scores of 77.0 for binary and 54.9 for multiclass classification, which can partly be attributed to the complex structure of the German language (Corazza et al., 2020). Outside of these challenges, the thus-far best-performing BERT-based multilingual classifier reaches an F1 score of 74.7 for German and 76.9 for French (Deshpande et al., 2022).

Despite this proliferation of research on hate speech, the topic has, to date, received very limited scholarly attention in the Swiss context, a particularly complex linguistic, sociocultural and political landscape. A notable exception is the work by Binder et al. (2020), who explore gendered hate speech in Swiss WhatsApp messages.

Research on post-deployment performance, classifier maintenance and infrastructural limitations is similarly scarce. Some attempts at discussing post-deployment performance in terms of cross-domain and concept drifts have been made (Vidgen et al., 2019; Sood et al., 2012; Aljero and Dimililer, 2021; Zhang et al., 2018; Malik et al., 2022; Ribeiro et al., 2020; Salminen et al., 2020; Bosco et al., 2018; Karan and Šnajder, 2018) and show that performances of deployed models are often lower than those obtained *in vitro*, sometimes dropping as much as 15 F1 points (Arango et al., 2019). While these insights are valuable, all of these studies only artificially post-test models with publicly available datasets: none of them presents long-term pre- and post-deployment results.

Similarly, infrastructural challenges encountered in the process of training, evaluating and testing remain understudied. Salminen et al. (2021) provide the currently most complete examination of challenges related to automatic hate speech detection. Of the 19 action points Salminen et al. describe, the setup of our classification pipeline allows us to contribute to addressing the following five: AP02 (developing a multilingual hate speech classifier), AP03 (acquiring new samples periodically to refresh the training sets used for model development), AP13 (collecting and testing datasets from multiple social media platforms), AP18 (measuring and displaying overall error rates of online hate classifiers to end-users such as moderators) and AP19 (providing probabilities and explanations of labels provided by the hate classifier for individual samples).

3 Methodology

3.1 Definition of Hate Speech

For this study, we opted for a combination of top-down and bottom-up approaches to establishing an onion-like definition of hate speech. At its heart lies the UN statement that hate speech is "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" (United Nations, 2019, p. 2). Additionally, we identified further characteristics and groups which were often targeted in online discourse in Switzerland. We therefore expanded our definition of hate speech to reflect those. Our final operational definition of hate speech comprises attacks and insults against the following target groups:

- sex, age, gender, religion, nationality/skin colour/origin, and mental and bodily impairments (UN definition)

- social status (e.g. income, education, job), political orientation and appearance

- other (e.g. Covid-19, cyberbullying)

In practice, it is not straightforward to delimit toxic speech, i.e. insults and derogatory use of language, from statements targeting a specific person or group based on their (inherent) characteristics. Many toxic claims linguistically closely resemble hate speech statements. Therefore, instances of toxic speech were also annotated as positive examples. In so doing, we could systematically evaluate how our classifier configured for hate speech performs in comparison to one fine-tuned on a looser, toxicity-based definition which corresponds more closely to many of the classification targets used in research to date.

3.2 Human-in-the-loop Hate Speech Classification Pipeline

Our objective was to build a binary, multilingual, human-centered, long-term sustainable and self-sufficient classification pipeline which enables robust identification of (online) hate speech and allows our NGO and media partners to react to hate speech in a more timely fashion. So far, classification pipelines often end after prediction has taken

place (Abro et al., 2020; Aljero and Dimililer, 2021; Mullah and Zainon, 2021; Pereira-Kohatsu et al., 2019). While this is reasonable in research settings, it does not allow the pipeline to remain up-to-date long-term, given the constant flux in language usage, style, discourse and topics of discussion (Florio et al., 2021). To bypass this issue, we propose incorporating an automatic retraining process into the pipeline, similar to that suggested by Alsafari and Sadaoui (2021), with the exception that our retraining involves a human-machine hybrid decision making process as part of the active learning and refining framework stages (Budd et al., 2019; Enarsson et al., 2022).

In this setup, two sets of labels are produced: weak labels, which are those generated by the classifier prior to human checks, and strong labels, which are those our annotators confirmed to be (non-)hate speech. In practice this means that an initially trained classifier is used to classify new data. The resulting weak labels are then checked by human coders who confirm or reject them and annotate the target groups. Once checked, the annotated examples – drawn from newspaper comments or tweets – are added to the existing training set before the model is automatically retrained either after a certain period of time has passed or a certain volume of new content is available. The progressive accumulation of annotated data and knowledge about hate speech allows the model to learn incrementally and adapt to changes, thus yielding more robust long-term results. To ensure that the model does not overfit and that parameters remain up-to-date, checks of the training results are carried out with parameter tuning taking place when and where necessary. The latter can be facilitated by a built-in function which, when called, launches a define-by-run API provided by the open-source optimisation software OPTUNA (Akiba et al., 2019) and automatically searches for the best hyperparameters using a search-and-prune algorithm for cost-effective optimisation.

3.3 Data Set and Data Set Balancing

All datasets stem from the Swiss Hate Speech Corpus specifically compiled for this study. The corpus, which is described in Appendix A in more detail, was annotated by a total of 19 trained student research assistants as well as volunteers recruited through our NGO partner. Annotators were instructed first to decide whether or not a given ex-

ample features hate speech or toxic speech. In cases where an example was hate speech, annotators were instructed to carry out multilabel annotation of the groups targeted. In cases where no target group was recognisable, annotators were asked to label the example as toxic speech. Most examples were annotated only a single time. Quality control was ensured by measuring the inter-rater-reliability for a subset of examples which were included in each annotation set, by performing qualitative checks and by holding bi-weekly deliberation rounds with annotators to clarify unclear cases.

The latter was particularly important given both the difficulty of the task and the heterogeneous nature of the data. While we focused on examples written in (Swiss) German and French, our data also includes numerous examples which feature terms in other national and frequently used languages in Switzerland, different Swiss German dialects as well as various writing scripts (e.g. Arabic, Cyrillic, Greek) and reading directions (e.g. Arabic). All of this introduced a host of linguistic and graphical issues which needed to be accounted for during the classification process. While other studies, like Ousidhoum et al.'s (2019), have excluded examples featuring code switching from their training sets, this was not an option for us given the inherently multilingual and multi-dialectal context of Switzerland.

With its over 422k unique annotations, this corpus is therefore one of the largest and linguistically most diverse corpora on hate speech (Poletto et al., 2021). The types and sources of data featuring in it were chosen in accordance with the requirements set by our NGO and media partners, without whom we would not have been able to carry out our project. As such, the corpus includes data from the following three main sources: (1) comments posted under online newspaper articles collected by our NGO partner (NGO); (2) online newspaper comments directly obtained from three Swiss national online newspaper outlets (ON1, ON2 and ON3); (3) tweets collected using the Twitter API of "politically interested users", i.e. accounts following at least five Swiss newspapers or politicians.

For our purposes, the most important set was the second, as it includes published, moderated and deleted comments from the three ONs, whose commentators – taken together – display varying linguistic, cultural and sociodemographic characteristics and are located all over Switzerland. To configure the classifier, we used all donated data, regardless of their moderation status. As expected from a corpus annotated for hate speech and the corresponding target groups (cf. Tita and Zubiaga, 2021), there are considerable class imbalances within the dataset (Geschke et al., 2019).

To mitigate these class imbalances and minimise overfitting, in the early stages of our work, we tested SMOTE, and numerous over- and undersampling strategies as well as different balancing ratios. Despite the distributional differences in the training and prediction sets, the best out-of-sample results were obtained by undersampling the majority class so as to obtain a 50-50 ratio of positive and negative examples in the training set. For the initial training set, all positively annotated instances were therefore joined with the same number of randomly sampled negatively annotated data points. For each retraining thereafter, the newly annotated data was balanced in such a way as to include an equal number of positive and negative examples, and was then added to the existing training set to produce an incrementally growing body of annotated examples. The same procedure was followed after deployment.

3.4 Classification Models

For training the classifier, we relied on two main architectures: multinomial Naïve Bayes (MNB), given its capability of performing text classification with few data points and in a computationally efficient manner, and on mBERT($_{Base}$ uncased) (Devlin et al., 2019; Pires et al., 2019; Aluru et al., 2020), the thus-far most promising model for performing classification tasks in non-English languages due to its bidirectional nature and pre-training on 104 languages. We also fine-tuned and tested a series of other BERT models using the HuggingFace open source library (Wolf et al., 2019): GELECTRA$_{Base}$ and GBERT$_{Base}$ (Chan et al., 2020), FlauBERT (Le et al., 2019), French RoBERTa$_{Base}$ (Majumder, 2021), XLM-RoBERTa$_{Base}$ (Conneau et al., 2019) and Twitter-XLM-RoBERTa$_{Base}$ (Barbieri et al., 2021). These models were chosen because they feature a wide variety of writing and language styles in their pre-training, which is important given the heterogeneous composition of our corpus; they also yield the most promising benchmark results for our type of task.

3.5 Preprocessing

Preprocessing steps for the MNB included the following: regularisation and stripping of extra white spaces, line breaks and multiple quotation marks, case folding, removal of punctuations, numbers, special characters, htmls and @-mentions, transcription of emojis to words, tokenisation, stopwords removal, lemmatisation. For the BERT-models, which are capable of grasping the semantics of sentences and taking into account word order, minimal preprocessing was carried out: regularisation, stripping of extra white spaces and line breaks, case folding, removal of htmls and @-mentions, transcription of emojis to words. For all steps, we used standard state-of-the-art libraries (nltk, re, emoji).

Transcribing emojis to words always means transcribing into English no matter the language of the rest of the input, which speaks in favour of using multilingual models. Moreover, we removed @-mentions given that our dataset displays some user names like @*schwurbler* (slang for conspiracy theorist) which themselves are derogatory terms. Removing them helps to prevent our model from classifying potentially harmless examples as hate speech based simply on the user name.

3.6 Training Parameters and Infrastructure

The bag-of-words MNB model was paired with tfidf and n-grams to ensure that at least some word order was preserved. A series of tests showed that best results are achieved with a maximum of 3000 tfidf features paired with a word n-gram range of 1-4 grams. The latter is based on Malmasi and Zampieri (2017) and Jauhiainen et al. (2018), who showed that 4-grams are most effective when processing (Swiss) German. Since our dataset also features numerous shorter words, including 1-3-grams further improved results. Model performance was evaluated using 10-fold-cross validation.

All transformer models were fine-tuned using the original model hyperparameters in order to allow for better comparisons between their performances. Leveraging data parallelisation, the models were trained for up to 5 epochs with a per-device batch size of 8 on a node consisting of 16 NVIDIA Tesla K80 GPUs with 16GB of system memory per GPU, running on average for approximately eight hours per training. The small batch size was conditioned on the fact that model parallelisation was not possible within our infrastructure, leading to out-of-

memory (OOM) if a larger batch size was chosen. The latter constraint also meant that most attempts to perform hyperparameter tuning led to overfitting. Saving the best model ensured that the best performance was then used for final evaluation and prediction. All post-deployment performance evaluations were carried out on a single node consisting of 8 NVIDIA Tesla V100 GPUs with 32GB of system memory per GPU and a per-device batch size of 32, taking on average approximately four hours per prediction task. All GPUs ran with cuda 11.2 and cudnn 8.1.0. Models were trained and evaluated using a stratified, random train-test split with an 80-20 ratio.

3.7 Deployment

For deployment, the classifier was integrated into a larger pipeline where new data in form of comments and tweets are periodically ingested from different Swiss online newspaper outlets and the Twitter API respectively. They are then passed through our human-in-the-loop hate speech classification pipeline which can be accessed by the NGO community and content moderators via a custom-built app hosted on university servers. In addition to presenting the comments and tweets in descending order of hate speech probability, the app allows users to perform checks of the hate speech labels and to carry out counter speech. Since all classification results are saved in our database, checked comments and tweets are automatically balanced and added to the training set for the next retraining round of the model.

4 Results and Discussion

To train and evaluate our classifier, we performed a series of experiments which align with the three stages of our development process: (1) incremental training which took place during the annotation process and where the classifier was retrained on a weekly basis; (2) pre-deployment experiments to cross-examine the final model results; and (3) post-deployment experiments where we tested the best-performing model *in situ*. With the exception of one pre-deployment experiment, all pre-deployment training was carried out on data where occurrences of toxic speech were treated as positive examples.

All experiments are reported using the weighted F1 score to account for the fact that post-deployment evaluation took place on a random sub-

		Train & Eval			Pred		MNB			mBERT$_{Base}$		
Wave	Week	Lang	Data	Size	Lang	Data	Precision	Recall	F1	Precision	Recall	F1
1	1	G	ON1	2.3k	G	ON1	62.2	62.3	62.2	OF	OF	OF
	2	G	ON1	4.1k	G	ON1	65.2	65.2	65.2	OF	OF	OF
	3	G	ON1	5.5k	G	ON1	63.2	63.1	63.0	OF	OF	OF
	4	G	ON1	8.7k	G	ON1	65.0	64.9	64.9	OF	OF	OF
	5	G	ON1	12.1k	G	ON1	64.0	64.0	64.0	OF	OF	OF
	6	G	ON1	14.5k	G	ON1	67.4	67.4	67.3	75.9	75.8	75.8
	7	G	ON1	20.5k	G	ON1	65.4	65.3	65.3	73.6	73.3	73.2
	8	G	ON1	28.4k	G	Tweets	66.6	66.5	66.4	73.3	72.7	72.5
	9	G	ON1 & Tweets	37.3k	G	Tweets	67.1	66.9	66.8	72.3	72.3	72.2
2	1	G	ON1	40.8k	FR	ON1	65.9	65.1	64.6	73.6	73.6	73.6
	2	G & FR	ON1	47.6k	FR	ON1	73.0	69.2	67.8	77.7	77.5	77.4
	3	G & FR	ON1	60k	FR	ON1	73.3	70.9	70.1	77.4	77.4	77.4
	4	G & FR	ON1	67k	FR	ON1	71.9	70.4	69.8	77.3	77.1	77.1
	5	G & FR	ON1	82.5k	FR	ON1	74.7	73.9	73.7	79.0	79.0	79.0
	6	G & FR	ON1	94.5k	FR	ON1	75.2	74.8	74.7	79.3	79.3	79.2
	7	G & FR	ON1	102.5k	FR	ON1	75.7	75.5	75.4	79.3	79.2	79.2
	8	G & FR	ON1 & Tweets	119k	FR	Tweets	74.2	73.8	73.7	79.1	79.0	79.0
	9	G & FR	ON1 & Tweets	122k	FR	Tweets	72.7	72.5	72.4	78.6	78.4	78.4
3	1	G & FR	ON1 & Tweets	127.7k	G	Tweets	71.5	71.4	71.4	78.4	78.4	78.4
4	1-15	G & FR	ON1 & Tweets	144.6k	G	Tweets	69.4	69.2	69.1	79.3	79.2	79.2
5	1-3	G & FR	ON1 & Tweets	177.9k	G	ON1	68.4	67.5	67.2	78.8	78.8	78.8
	4-6	G & FR	ON1 & Tweets	186.7k	G	ON2	68.7	68.3	68.1	80.5	80.3	80.3
	7-9	G & FR	ON1-2 & Tweets	191.6k	G	ON3	68.7	68.3	68.1	80.6	80.3	80.3
	10-12	G & FR	ON1-3 & Tweets	197.4k	G	ONs & Tweets	69.3	68.8	68.6	80.8	80.5	80.5

Table 1: Incremental Training Results

set of human-checked examples which the system had classified as hate speech. As such, in addition to human checks, F1 remained the most important performance metric throughout this study. Without any labeled hate speech data or baselines against which new results could be judged, for each re-training, a new F1 was generated which, depending on the result, may or may not have become the new baseline. In other words, the baseline against which we judged new results was always the best result which the model had achieved in the previous training round(s). This human-in-the-loop active learning approach proved to be a successful strategy as it allowed us concurrently to generate labeled data and configure the classifier. The final result reported in this study is that obtained upon the completion of our project. While we hope that it can serve as the baseline for future (Swiss) German and French hate speech classifiers, (re)training could have, in theory, continued until the F1 scores and the number of human-checked hate speech stagnate.

Even though such a saturation point may eventually be reached *in vitro*, it is unlikely that this will ever happen *in situ*, given the shifting nature of language usage, topics of discussion and similar factors. Post-deployment retraining is therefore never-ending and is evaluated in the same way as during the annotation process. As such, F1 remains vital for post-deployment classifier maintenance. However, since the model is, upon deployment, fine-tuned for the (Swiss) German and French contexts, retraining would require fewer

new data points – just enough to introduce new features – and would not have to be carried out as frequently as we needed to perform it, but certainly in moments of 'dramatic' language shifts. As such, F1 can also be used for adjusting the retraining strategy and thus lowering carbon footprints, costs and not least human effort.

4.1 Phase 1: Incremental Training

The first experiment aimed at systematically tracing how model performances change from one retraining to the next and at identifying potential pitfalls occurring during incremental training. The results obtained are listed in Table 1, where "G" and "FR" indicate German and French, the source language of the data, respectively. The different data incrementation sizes reflect the number of annotated hate speech in any particular week and wave, and depended on the given stage of the annotation and retraining process, which model was used to generate weak labels, and on which language and data type we trained.

Unlike the MNB, mBERT could not be used for (re-)training early on in the process. Only once a training set size of 14.5k data points was reached, was it possible to fine-tune the model without it overfitting (OF). This can partly be traced to the sources of examples and their respective distributions within the dataset, to the linguistic and graphical complexity of the training set and to the fact that the majority of examples were annotated only a single time. The latter in particular means the model has a much harder time performing its task,

as was also suggested by Leonardelli et al. (2021). They argue that the higher the per-sample inter-inter-rater reliability, the fewer samples are needed for the model to achieve good overall performance.

At the moment of introduction, mBERT instantly outperformed and continued to outperform the MNB, as is also visible in the annotation results (cf. Table 10 in Appendix B) which see a tripling in identified hate speech comments. In the process of retraining, mBERT, save for some minor oscillations, improved continually, confirming Aluru et al.'s (2020) observation that performances of classifiers improve with increasing number of data points on which the models are trained. The most pronounced oscillations happened at the point where prediction took place on French tweets: both the MNB and mBERT dropped in performance, but while mBERT was able to recover after a few retrainings, MNB performance continued to drop.

A different picture arose during zero-shot cross-lingual classification in wave 2 week 1, when the models – fine-tuned on (Swiss) German comments only – were used to make predictions on (Swiss) French comments which up to this point had not featured in the training set. Rather than causing a drop in performance, as was observed by e.g. Pelicon et al. (2021), transfer learning between the two languages improved the F1 score by 1.4 points. Throughout the remainder of the iterative training, mBERT performance continued to improve, as is also reflected in the annotation results. For the MNB, the highest F1 score of 75.7 is recorded in wave 2 week 7; mBERT peaked in wave 5 week 12 with an F1 score of 80.5 and outperforms the thus-far best-performing BERT-based multilingual hate speech classifier (Deshpande et al., 2022) by 5.8 F1 points in German and 3.6 F1 points in French.

Data	Toxic	Precision	Recall	F1
G & FR Comments & Tweets	toxic as HS	79.3	79.2	79.2
G & FR Comments	no toxic	80.0	79.9	79.8
G & FR Comments	toxic as nonHS	72.5	71.9	71.8

Table 2: Model Performance according to Toxic Speech in the Training Set

4.2 Phase 2: Pre-Deployment

The second experiment aimed at establishing whether mBERT was indeed the best BERT model for our purposes. For this, we trained and evaluated a series of other models whose performance results are detailed in Table 11 in Appendix B. For each model and dataset, the transformer model in question was first used for prediction (base) before it was fine-tuned and evaluated on our annotated dataset (tuned). Despite expecting, and getting, very poor results during the base runs, it emerged that model performance already at this stage is language and data type dependent. On average, GELECTRA performed best for German comments, and German comments and tweets, GBERT for German tweets, flauBERT for all French data types, XLM-RoBERTa and Twitter-XLM-RoBERTa for multilingual comments and XLM-RoBERTa for multilingual tweets.

This is a valuable observation given that implicit biases are present in these models and remain present even after fine-tuning. We noticed this propagated bias in wave 1 week 8 of the annotation process when the majority of tweets in our dataset still stemmed from Germany – we later adapted our filtering process to minimise this phenomenon. Even though our classifier was capable of picking up some of the hate speech tweets with a distinctly Swiss German context, mBERT was much more ready and capable to pick up hate speech in tweets written in German German. This is presumably because most of the German speaking Wikipedia texts on which the model was pretrained will have stemmed from a German context.

Once fine-tuned, the models yielded varying performances. Models fine-tuned and evaluated exclusively on (Swiss) German comments yielded similar results, whereby the monolingual outperformed the multilingual models. In French, flauBERT outperformed french RoBERTa, presumably because of the varied nature of the datasets on which it was pretrained. No results can be reported on French tweets given the small size of the training set which led to almost instantaneous overfitting. For all data types, the multilingual models performed better on the pure French sets than did the monolingual models, suggesting that monolingual models for French are not yet as accurate as their German counterparts. While the multilingual models performed similarly when trained and evaluated on multilingual datasets, mBERT significantly outperformed XLM-RoBERTa and Twitter-XLM-RoBERTa once both comments and tweets were used. We therefore proceeded to deploy this model.

However, before final deployment, we carried out another experiment to examine whether or not and, if so, how toxic speech should be incorporated

Train & Eval Data	Train & Eval Platform	Test Platform	Eval			Test		
			Precision	Recall	F1	Precision	Recall	F1
G & FR Comments	ON1	ON1	79.8	79.6	79.6	78.8	78.8	78.8
G & FR Comments	ON1	Twitter	79.8	79.6	79.6	66.3	71.6	65.4
G & FR Tweets	Twitter	ON1	76.4	76.4	76.4	68.7	68.1	68.3
G & FR Tweets	Twitter	Twitter	76.4	76.4	76.4	70.7	70.1	70.7
G & FR Comments & Tweets	ON1 & Twitter	ON1	79.3	79.2	79.2	75.8	74.3	74.9
G & FR Comments & Tweets	ON1 & Twitter	Twitter	79.3	79.2	79.2	70.6	70.1	69.4

Table 3: Cross-Platform Classification Results

Data	Train & Eval Data	Test Data	Eval			Test		
			Precision	Recall	F1	Precision	Recall	F1
G & FR Comments	ON1	ON1	79.8	79.6	79.6	78.8	78.8	78.8
G & FR Comments	ON1	ON2	79.8	79.6	79.6	77.9	82.3	75.5
G & FR Comments	ON1	ON3	79.8	79.6	79.6	92.2	96.0	94.1
G & FR Comments	ON1 + ON2	ON3	80.6	80.3	80.3	92.2	96.0	94.1
G & FR Comments	ON1 + ON3	ON2	80.8	80.7	80.6	75.0	82.0	75.0

Table 4: Cross-Dataset Classification Results

in the training set. Throughout the annotation process, occurrences of toxic speech were annotated as positive examples. Since our classifier is meant primarily to look for hate speech not toxic speech in the wider sense, three tests were carried out: in the first, examples featuring toxic speech remained under the positive label, as they did in all other experiments; in the second, they were dropped from the training set; in the third, examples of toxic speech were included in the training set as negative examples.

The results of these experiments are summarised in Table 2. They show that reclassifying toxic speech to the negative label significantly decreased performance. Leaving toxic speech under the positive label or dropping toxic speech entirely yielded similar quantitative performances. Thereby the classifier without toxic speech performed slightly better. A qualitative analysis confirmed that training the classifier without toxic speech sharpens the focus, leading to fewer toxic examples and more hate speech proper, i.e. attacks against specific target groups, being recognised.

Determining these differences was only possible by inspecting the predictions and manually comparing their results. The evaluation metrics only marginally helped, suggesting that in the future, in settings such as ours, evaluation metrics are needed which also account for qualitative performance differences. Given these results, the deployed mBERT classifier does not feature toxic speech in the training set.

4.3 Phase 3: Post-Deployment

To test the robustness and long-term sustainability of the classifier, further experiments were car-

ried out post deployment. Since post-deployment evaluation is identical to the active learning and incremental pre-deployment evaluation which took place during the annotation process, all annotations resulting from these tests were added to the Swiss Hate Speech Corpus (wave 5). This manner of proceeding allowed for further tracing of the incremental training results post-deployment and served as a stress-test for the hate speech classification pipeline when used *in situ*.

In the first experiment, we trained and evaluated the model on German and French ON1 comments, then on tweets and finally on ON1 comments and tweets, all taken from the first four annotation waves, before using the deployed mBERT and the MNB for comparison to predict on data from ON1 and Twitter. As the results in Table 3 show, zero-shot cross-platform classification significantly decreased model performance with training on ON1 and prediction on Twitter causing the largest drop of 14.2 F1 points.

Similar results emerged from the second experiment in which we performed cross-dataset testing on ON1, ON2 and ON3. To ensure the model would not overfit due to small numbers of training samples, ON1 comments annotated throughout the first four waves featured in each of the training sets. Prediction was performed on a random sample of the ON which the model had not yet seen. As Table 4 shows, performance dropped for all ONs up to 5.6 F1 points except for ON3 where we obtained significantly higher results than during evaluation. This can be explained by the fact that ON3, unlike ONs 1 and 2, features clearer cases of hate speech.

In the final experiment, we tested for temporal drift by using different temporal cut-offs to create

			Eval			Test		
Train Data	Train and Eval Period	Testing Period	Precision	Recall	F1	Precision	Recall	F1
G & FR Comments	01-2019 to 06-2019	07-2019 to 07-2021	74.6	74.5	74.5	78.2	51.5	51.5
G & FR Comments	01-2019 to 12-2019	01-2020 to 07-2021	80.1	79.6	79.5	58.9	62.5	60.0
G & FR Comments	01-2019 to 06-2020	07-2020 to 07-2021	80.0	79.6	79.5	83.9	75.8	79.0
G & FR Comments	01-2019 to 12-2020	01-2021 to 07-2021	79.9	79.7	79.6	81.6	74.6	77.2
G & FR Comments	01-2019 to 06-2021	07-2021 to 07-2021	79.2	79.1	79.1	82.1	76.8	78.9
G & FR Comments	01-2019 to 07-2021	01-2022 to 03-2022	80.0	79.9	79.8	65.1	71.5	65.5

Table 5: Temporal Drift

HS Probability	Total	HS	nonHS	HS%
1.00-0.90	3995	3601	394	90
0.89-0.85	3952	2946	1006	75
0.89-0.80	7717	5458	2259	71
0.79-0.70	7952	4197	3755	53
0.69-0.60	7951	3274	4677	41
0.59-0.50	6207	1897	4310	31

Table 6: Thresholding

various combinations of training and test sets. The results are summarised in Table 5 and suggest that the more offset in time the prediction set is, the less accurate the classifier becomes. This is particularly visible in the last test where the training and prediction sets are almost 6 months apart. The result of this time shift was a drop in performance by 14.3 F1 points.

To mitigate some of this classifier drift and provide more robustness to the deployed model, we conducted an analysis of how hate speech samples behave according to predicted hate speech probability. In practice, this allows us to provide empirically-grounded recommendations for which hate speech threshold end users should use when deploying our model. As shown in Table 6, the number of hate speech examples decreases the further down the probability scale we go. This was also reflected in qualitative analyses which showed that the further down the probability scale we go, the more contested and borderline cases we find.

Striking in our case is the sharp decline in the number of hate speech tweets in the 0.89-0.80 probabilities band when compared to that between 1.00 to 0.90. For our deployed model, this indicates that if thresholding were to be applied, the optimal cutoff point would be at 0.90. Depending on how much tolerance is acceptable, the threshold could be lowered: yet to guarantee the model performance obtained at the time of training and evaluation, the threshold should be set no lower than 0.85, at which point 80 percent of positively classified examples will still be hate speech. Any threshold below 0.85 would no longer guarantee optimal model performance.

5 Conclusion

This paper presents a multilingual human-in-the-loop BERT-based hate speech classification pipeline adapted to the Swiss context. By systematically tracing and reporting classifier performances from the start of the annotation process to post-deployment, we highlight important practical considerations for fine-tuning a state-of-the-art BERT-based classification model. The most important is that testing classifiers post-deployment is vital for configuring robust systems, as various types of drift cause performances to drop. Next to focusing on ameliorating systems pre-deployment, future research should thus also look into possible mitigation techniques to stabilise classifier performance post-deployment.

In addition to recommending thresholds, a first step in that direction is to discard the idea of a static, one-time-only-trained model in favour of an adaptive learning concept (Gama et al., 2014; Laaksonen et al., 2020). In this paradigm, the classifier functions like a living organism and evolves together with the ever-changing uses of language, and alterations in mindsets and topics. This can only be achieved by actively maintaining the classifier post-deployment, i.e. by dynamically adjusting the definition of hate speech and the annotation schemes, by periodically retraining the model and by allowing it progressively to accumulate and adapt its knowledge of hate speech.

6 Limitations

The biggest issues we faced throughout this study were infrastructural limitations related to scalability. Due to NDAs preventing us to upload our raw newspaper comment data to cloud services and similar infrastructures, all training, evaluation and testing was carried out on the university cluster. With each incremental increase in training set size, more memory and time was needed for retraining the model. Restrictions on the availability of GPU memory in particular meant that, despite leveraging

data parallelisation, we could not train above a per-device batch size of 8 which meant that the number of requested GPUs had to be increased every couple of rounds in order to maintain manageable training and prediction times. The latter was particularly important since queuing times on the cluster could range anywhere from no waiting to waiting up to a week, potentially causing significant delays, particularly during the annotation process when the classifier had to be retrained weekly.

To increase the batch size and try further hyperparameter tuning, we made several attempts at parallelising the model using Microsoft's (2022a) DeepSpeed optimisation library. The problem we encountered was that at the time of implementation, DeepSpeed appeared only to be configured for single user multi-GPU parallelisation which led to significant disruptions when implemented on a shared-user cluster. The behaviour we observed was that DeepSpeed reindexes available GPUs rather than inheriting the GPU numbers of the cluster, meaning that GPUs were called which were, in fact, not available, causing disruptions in other users' work. Despite trying a series of work around options in the form of intercepting the launch of DeepSpeed at an earlier point and avoiding this reassignment, we were unable to solve the problem from our end. We subsequently reported this issue to Microsoft but have not yet received a reply.

Exploring options of how to achieve better performance using zero- and few-shot models (Brown et al., 2020) as well as more powerful x-former models operating with more computational and memory efficiency (Tay et al., 2020) is helpful. However, given our experience, it would also be desirable if future research shifted the focus more towards developing further options to parallelise models. Ideally, this would be done in a manner where software companies work together with research institutions in order to ensure better compatibility of available libraries and infrastructures. Another desirable direction of research would be for software companies to find a way to configure their cloud services in such a way that using them would not breach data protection requirements which do not allow uploading raw data to external parties.

A further infrastructural issue was encountered during deployment: at the moment, our hate speech classification API is running on university servers where we can offer the same infrastructure as was used during the configuration process and can thus provide an estimate of training and prediction times as well as costs. We are also still in charge of periodically checking classifier performance once the model is retrained. However, once our NGO partner and media companies decide to continue using the classifier without our input, they may no longer have the corresponding infrastructural set-up to run and maintain the classifier. Future work hoping to make contributions to both research and industry would thus benefit from implementing infrastructural frameworks which are tenable for industry partners. While research, like the work by Tsankov et al. (2022), shows that this field is slowly gaining in importance, more needs to be done to facilitate easy deployment and long-term maintenance of AI systems.

7 Ethics Statement

While our dataset is more diverse than other hate speech benchmark sets, it is also prone to bias, a problem already addressed in numerous studies (Binns et al., 2017; Bender and Friedman, 2018; Gebru et al., 2018; Holland et al., 2018; Davidson et al., 2019; Ranasinghe et al., 2019; Gorwa et al., 2020; Kim, 2021) advocating for more transparency when it comes to constructing corpora from naturally occurring data. In our case, bias is introduced on four main levels: by the sources and platforms, and their respective representation in the dataset, by the manner in which training data were annotated throughout the annotation process, by the pre-trained model and by the prevalence of hate speech and distribution of target groups in the Swiss Hate Speech Corpus.

7.1 Bias Stemming from Sources and Platforms

Since our training set is mainly composed of data points from ON1 and Twitter, the classifier is well-tuned to examples stemming from these sources. When they occur during production, prediction is likely to achieve the performance results obtained during evaluation. However, as the cross-dataset and platform experiments show, the model does not yet generalise particularly well, which is partly conditioned on the fact that sources and platforms are represented differently in the dataset. ON2 and ON3, for instance, were only accessible to us towards the end of the study for final *in situ* testing. Accordingly, a minority of the dataset covers comments from these two sources. Similarly, French

tweets are – compared to (Swiss) German tweets – underrepresented in the corpus. While transfer learning certainly ensures that the classifier still achieves high performance, a more balanced representation of the different sources could lead to more model robustness.

At the same time, more generalisibility could be achieved by including into the training set examples from other online newspaper outlets and media platforms, like Facebook, Reddit, Signal or WhatsApp. Access to comment (or post) data from these platforms varies but since our classifier is periodically being retrained, these sources could easily be added in the future. The same applies for languages: our classification pipeline is based on mBERT, so if, in the future, data points in Italian, Romansh or other languages were to become available, they could easily be integrated into the existing dataset and – through transfer learning – potentially and naturally help to de-bias the set and enhance the performance. This approach is especially promising given the fact that French, Italian and Romansh belong to the same language family (Deshpande et al., 2022).

7.2 Annotation Bias

Defining hate speech is difficult in itself, but recognising it in real-life examples is even more difficult because the phenomenon keeps shifting and is entirely based on the concept of language usage and, thus, the context in which it was posted. The latter is not always available or clear enough, so decisions need to be made according to one's own best judgement. Existing research suggests that these judgments often depend on a series of annotator characteristics, like rational, background, linguistic competence and understanding of (the level of) hate speech (Wiegand et al., 2019; Mathew et al., 2020; Ousidhoum et al., 2020; Huang et al., 2020; Asogwa and Onwuama, 2021).

In our case, the majority of annotators were university students, most of them in political science, and aged 20-27. With similar educational backgrounds and age groups, it is likely that annotations in the training set reflect how they interpret hate speech, especially since the majority of the comments and tweets in our training set were annotated by only a single annotator; for even with the guidelines and support we provided to them, the underlying understanding of when language usage crosses into hate speech is ultimately theirs.

Furthermore, as outlined in Appendix A in more detail, our annotators themselves reported difficulties when annotating comments and tweets, which will have further influenced their judgements.

Quality controls were carried out to ensure coherence between the different annotation sets, but due to limitations in resources only a limited number were annotated a second or even a third time. Common themes we noticed throughout our deliberation rounds as well as during quality control and second annotations were that our annotators were particularly struggling to separate hate speech from various types of humour, enraged personal opinions, and opinions and feelings which do not reflect their own. In terms of humour, irony (Gibbs Jr., 2021), sarcasm (Teh et al., 2018) and euphemisms (Ayunts and Paronyan, 2021) were often the sources of concern given their provocative nature and the difficulties associated in identifying them in the written, sometimes context-less, medium, where neither paraverbal communication nor body language can be used as indicators. Therefore, a more in-depth analysis of the annotated dataset would need to be carried out in the future to determine the extent to which our annotated dataset is biased towards the opinions of our research assistants.

In the meantime, volunteers working with our NGO partner as well as moderators who are currently using the deployed classification pipeline will, as central actors in the pipeline and with time, progressively de-bias or at least shift the bias of the system. If, as we envisage it, the network of our users can be extended to include NGOs and moderators from the Italian and Romansh speaking parts of Switzerland, of neighbouring countries and possibly even the Netherlands, Luxembourg and Belgium given the linguistic similarities between the respective languages and the issues they face with language- and context insensitive classification systems, even more diversification and further de-biasing of the model might be possible.

7.3 Model Bias

While dataset bias has received most attention to-date, model bias is just as important an issue in configuring a hate speech detection system. As described in the results section, the pretrained models themselves carry various types of biases. We specifically encountered linguistic bias, likely introduced by the datasets on which the models were originally trained. In addition to fine-tuning the pre-

trained models, future research would benefit from placing more emphasis on de-biasing the models before fine-tuning them by following examples set by e.g. Bolukbasi et al. (2016); Basta et al. (2019), who have attempted to remove gender bias from word embeddings.

7.4 Bias Stemming from the Prevalence of Hate Speech and Target Groups

Part of the reason we decided to build a binary rather than a multilabel classifier was that the Swiss Hate Speech Corpus, as detailed in Appendix A, includes a huge disparity in numbers of samples per target group category. Since these are naturally occurring data points, they reflect the prevalence of hate speech in Switzerland and its corresponding distribution across different target groups, but they also pose significant issues in terms of data imbalance for training multilabel classifiers. Attempts at balancing out the dataset with the help of a series of common regularisation techniques (Srivastava et al., 2014; Kukacka et al., 2017; Venturott and Ciarelli, 2020; Feng et al., 2021; Venturott and Ciarelli, 2021) did not yield any useful results. Neither weighting classes, nor interpolating with SMOTE (Chawla et al., 2011; Fernández et al., 2018) worked.

Automatically augmenting the minority class with techniques like word splitting, syntax permutations, synonym replacement, insertions, deletions and substitutions as well as word2vec and contextual word embeddings with BERT, in order to upsample the minority class (Marivate and Sefara, 2019; D'Sa et al., 2021) proved that label preservation was not possible, with mutations of many hate speech samples yielding adversarial non-hate speech instances under the hate speech label. It also showed that most instances of augmented data preserved the form, but not the often socially attributed meaning; that in the case of Swiss German dialects not even the form could be preserved; that, as Shorten et al. (2021) have already remarked, the techniques work better on longer than on shorter input sequences, leading to many problematic statements in short comments and especially in tweets; and that automatic spell checking was not configured for such a large amount of spelling errors and could, in some cases, not even identify the language of a specific comment or tweet. Given the sensitive nature of the subject and studies like Longpre et al. (2020), who have shown that data augmentation

techniques do not improve the performance of pretrained models capable of processing syntax and semantics, we did not employ these augmentation techniques.

Instead, more advanced data balancing techniques as described in e.g. Mrksic et al. (2016); Shorten et al. (2021); Feng et al. (2021); Bayer et al. (2021) and allowing for label preservation (Alzantot et al., 2018; Kobayashi, 2018; Rizos et al., 2019) should be explored in the future. One of these methods, automatic back/round trip translation (Beddiar et al., 2021), has already been implemented on a subset of German tweets in a course paper by one of our annotators, yielding promising first results. If these techniques are successful or once more pertarget group samples become available from our deployed model, a multilabel classifier, be this as a single or ensemble model, could be envisaged which would not bias the model towards certain target groups while neglecting others.

References

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Zafar Ali, Sajid Khan, and Ghulam Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(August).

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A nextgeneration hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data*

Mining, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Mona Khalifa A. Aljero and Nazife Dimililer. 2021. A novel stacked ensemble for hate speech recognition. *Applied Sciences*, 11(24).

Safa Alsafari and Samira Sadaoui. 2021. Semi-supervised self-training of hate and offensive speech from social media. *Applied Artificial Intelligence*, 0(0):1–25.

Raghad Alshaalan and Hend Al-Khalifa. 2020. Hate speech detection in saudi twittersphere: A deep learning approach. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 12–23, Barcelona, Spain (Online). Association for Computational Linguistics.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *CoRR*, abs/1804.07998.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, New York, NY, USA. Association for Computing Machinery.

N.U. Asogwa and M.E. Onwuama. 2021. Hate speech and authentic personhood: Unveiling the truth. *SAGE Open*.

A. Ayunts and S. Paronyan. 2021. From euphemism to verbal aggression in british and armenian cultures: A cross-cultural pragmatic perspective. *FLEKS - Scandinavian Journal of Intercultural Theory and Practice*, 7(1):26–42.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Francesco Barbieri, Luis Espinosa Anke, and José Camacho-Collados. 2021. XLM-T: A multilingual language model toolkit for twitter. *CoRR*, abs/2104.12250.

Matthew Barnidge, Bumsoo Kim, Lindsey A. Sherrill, Žiga Luknar, and Jiehua Zhang. 2019. Perceived exposure to and avoidance of hate speech in various communication settings. *Telematics Informatics*, 44.

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Christine Basta, Marta Ruiz Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *CoRR*, abs/1904.08783.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *CoRR*, abs/2107.03158.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *CoRR*, abs/2106.04681.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Larissa Binder, Simone Ueberwasser, and Elisabeth Stark. 2020. Gendered hate speech in swiss WhatsApp messages. In *Language, Gender and Hate Speech A Multidisciplinary Approach*. Fondazione Università Ca' Foscari.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. *CoRR*, abs/1707.01477.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples [online].*, Torino. Accademia University Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish,

Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Beatrix Brügger, Rafael Lalive, and Josef Zweimüller. 2009. Does culture affect unemployment? evidence from the röstigraben. *IZA Institute of Labor Economics*, 4283:1–44.

Samuel Budd, Emma C. Robinson, and Bernhard Kainz. 2019. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *CoRR*, abs/1910.02923.

Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making. URL: https://orca.cardiff.ac.uk/65227/1/IPP2014-Burnap.pdf, access date 22.02.2022.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Prateek Chaudhry and Matthew Lease. 2020. You are what you tweet: Profiling users by past tweets to improve hate speech detection. *CoRR*, abs/2012.09090.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 351–360, Toulouse, France. ATALA.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A Multilingual Evaluation for Online Hate Speech Detection. *ACM Transactions on Internet Technology*, 20(2):1–22.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*.

Neha Deshpande, Nicholas Farris, and Vidhur Kumar. 2022. Highly generalizable models for multilingual hate speech detection. *CoRR*, abs/2201.11294.

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.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 29–30, New York, NY, USA. Association for Computing Machinery.

Armend Duzha, Cristiano Casadei, Michael Tosi, and Fabio Celli. 2021. Hate versus politics: Detection of hate against policy makers in italian tweets. *CoRR*, abs/2107.05357.

Aashwin Geet D'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. 2021. Exploring conditional language model based data augmentation approaches for hate speech classification. In K. Ekštein, F. Pártl, and M. Konopík, editors, *Text, Speech, and Dialogue. TSD 2021. Lecture Notes in Computer Science*, volume 12848. Springer.

Therese Enarsson, Lena Enqvist, and Markus Naarttijärvi. 2022. Approaching the human in the loop - legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, 31:1:123–153.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *CoRR*, abs/2107.08720.

Federal Department of Foreign Affairs FDFA. 2020. Four languages, four cultures – one country. URL: https://www.eda.admin.ch/eda/en/fdfa/fdfa/aktuell/newsuebersicht/2020/09/mehrsprachigkeit-1.html, access date 14.06.2022.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. *CoRR*, abs/2105.03075.

Alberto Fernández, Salvador García, Francisco Herrera, and Nitesh V. Chawla. 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary. *Journal of Artificial Intelligence Research*, 61(1):863–905.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*.

Komal Florio, Valerio Basile, and Viviana Patti. 2021. Hate speech and topic shift in the covid-19 public discourse on social media in italy. In *CLiC-it*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4).

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010.

Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. Classification of Hate Speech Using Deep Neural Networks. *Revue d'Information Scientifique & Technique*, 25(01).

Daniel Geschke, Anja Klaßen, Matthias Quent, and Christoph Richter. 2019. HASS IM NETZ: Der schleichende Angriff auf unsere Demokratie. Eine bundesweite repräsentative Untersuchung. *Verfügbar am*, 25:2020.

Raymond W. Gibbs Jr. 2021. "Holy Cow, My Irony Detector Just Exploded!" Calling Out Irony During The Coronavirus Pandemic. *Metaphor and Symbol*, 36(1):45–60.

Robert Gorwa, Reuben Binns, and Katzenbach Christian. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*.

Gunther M. Hega. 2001. Regional identity, language and education policy in switzerland. *Compare: A Journal of Comparative and International Education*, 31(2):205–227.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR*, abs/1805.03677.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective API built for detecting toxic comments. *CoRR*, abs/1702.08138.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *CoRR*, abs/2002.10361.

Sylvia Jaki and Tom De Smedt. 2019. Right-wing german hate speech on twitter: Analysis and automatic detection. *ArXiv*, abs/1910.07518.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018. HeLI-based experiments in Swiss German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 254–262, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. *CoRR*, abs/2104.14337.

Jae Yeon Kim. 2021. Power, Hate Speech, Machine Learning, and Intersectional Approach. SocArXiv chvgp, Center for Open Science.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Sujatha Arun Kokatnoor and Balachandran Krishnan. 2020. Twitter hate speech detection using stacked weighted ensemble (swe) model. In *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pages 87–92.

Jan Kukacka, Vladimir Golkov, and Daniel Cremers. 2017. Regularization for deep learning: A taxonomy. *CoRR*, abs/1710.10686.

Maximilian Kupi, Michael Bodnar, Nikolas Schmidt, and Carlos Eduardo Posada. 2021. dictnn: A dictionary-enhanced cnn approach for classifying hate speech on twitter. *ArXiv*, abs/2103.08780.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, page 1621–1622. AAAI Press.

Salla-Maaria Laaksonen, Jesse Haapoja, Teemu Kinnunen, Matti Nelimarkka, and Reeta Pöyhtäri. 2020. The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data*, 3(3):1–16.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *CoRR*, abs/1912.05372.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *CoRR*, abs/2109.13563.

Shayne Longpre, Yu Wang, and Christopher DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? *CoRR*, abs/2010.01764.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14:e0221152.

Abhilash Majumder. 2021. French RoBERTa. URL: https://huggingface.co/abhilash1910/french-roberta/blob/main/README.md, access date 02.03.2022.

Jitendra Singh Malik, Guansong Pang, and Anton van den Hengel. 2022. Deep learning for hate speech detection: a comparative study. *CoRR*, abs/2202.09517.

Shervin Malmasi and Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *CoRR*, abs/1803.05495.

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 Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2020. Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in indo-european languages. *CoRR*, abs/2108.05927.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in english and indo-aryan languages. *CoRR*, abs/2112.09301.

Vukosi Marivate and Tshephisho Sefara. 2019. Improving short text classification through global augmentation methods. *CoRR*, abs/1907.03752.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *CoRR*, abs/2012.10289.

Joshua Melton, Arunkumar Bagavathi, and Siddhartha Krishnan. 2020. Del-hate: A deep learning tunable ensemble for hate speech detection. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1015–1022.

Meta. 2022. Facebook hate speech definition. URL: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/, access date 23.02.2022.

Microsoft. 2022a. Deepspeed. URL: https://github.com/microsoft/deepspeed, access date 23.05.2022.

Microsoft. 2022b. Microsoft hate speech definition. URL: https://www.microsoft.com/en-us/concern/hatespeech, access date 23.02.2022.

Mr. Mohiyaddeen and Sifatullah Siddiqi. 2021. Automatic hate speech detection: a literature review. *SSRN*.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT

'17, page 85–94, New York, NY, USA. Association for Computing Machinery.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. *CoRR*, abs/1910.12574.

Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. Counter-fitting word vectors to linguistic constraints. *CoRR*, abs/1603.00892.

Sean Mueller and Paolo Dardanelli. 2014. Langue, culture politique et centralisation en suisse. *Revue internationale de politique comparée*, 21(4):83–104.

Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: A review. *IEEE Access*, 9:88364–88376.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? Making sense of language-specific BERT models. *CoRR*, abs/2003.02912.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. 2020. Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2532–2542, Online. Association for Computational Linguistics.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020. The refugee experience online: Surfacing positivity amidst hate. In Giuseppe De Giacomo, Alejandro Catala, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang, editors, *ECAI 2020: Frontiers in Artificial Intelligence and Applications*, pages 2925–2926. IOS Press.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. Automatic identification of misogyny in English and Italian Tweet at EVALITA 2018 with a multilingual hate lexicon. In *EVALITA@CLiC-it*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 30–34, Online. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors (Basel, Switzerland)*, 19.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 199–207. CEUR-WS.org.

Priya Rani, Shardul Suryawanshi, Koustava Goswami, Bharathi Raja Chakravarthi, Theodorus Fransen, and John Philip McCrae. 2020. A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 42–48, Marseille, France. European Language Resources Association (ELRA).

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*,

CIKM '19, page 991–1000, New York, NY, USA. Association for Computing Machinery.

Björn Ross, Michael Rist, Guillermo Carbonell, Ben Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Z. Margetts, and Janet B. Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *ACL/IJCNLP*.

Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *CoRR*, abs/1709.10159.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon Gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 12th International AAAI Conference on Web and Social Media, ICWSM 2018, pages 330–339. AAAI press. Publisher Copyright: Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 12th International AAAI Conference on Web and Social Media, ICWSM 2018 ; Conference date: 25-06-2018 Through 28-06-2018.

Joni Salminen, Maximilian Hopf, S. A. Chowdhury, Soon-Gyo Jung, Hind Almerekhi, and Bernard Jim Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34.

Joni Salminen, Maria Jose Linarez, Soon-gyo Jung, and Bernard J. Jansen. 2021. Online hate detection systems: Challenges and action points for developers, data scientists, and researchers. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*, pages 1–7.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Connor Shorten, Taghi Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data*, 8.

Alexandra A. Siegel. 2020. *Online Hate Speech*, SSRC Anxieties of Democracy, page 56–88. Cambridge University Press.

Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *The Journal of the Association for Information Science and Technology*, 63(2):270–285.

Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44(2):136–146.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Patrick Stevenson. 1990. Political culture and intergroup relations in plurilingual switzerland. *Journal of Multilingual & Multicultural Development*, 11(3):227–255.

Julia Maria Struss, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *KONVENS*.

Dewa Ayu Nadia Taradhita and I Ketut Gede Darma Putra. 2021. Hate speech classification in indonesian language tweets by using convolutional neural network. *Journal of ICT Research and Applications*, 14:225–239.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. *CoRR*, abs/2009.06732.

Phoey Lee Teh, Pei Boon Ooi, Nee Nee Chan, and Yee Kang Chuah. 2018. A comparative study of the effectiveness of sentiment tools and human coding in sarcasm detection. *Journal of Systems and Information Technology*, 20(3):358–374.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment in short strength detection informal text. *Journal of the Association for Information Science and Technology*, 61:2544–2558.

Teodor Tita and Arkaitz Zubiaga. 2021. Cross-lingual hate speech detection using transformer models. *CoRR*, abs/2111.00981.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126:157–179.

Petar Tsankov, Pavol Bielik, Martin Vechev, and Andreas Krause. 2022. LatticeFlow. URL: https://latticeflow.ai, access date 23.02.2022.

Twitter. 2022. Twitter hateful conduct definition. URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, access date 23.02.2022.

United Nations. 2019. United Nations Strategy and Plan of Action on Hate Speech. URL: https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf, access date 23.02.2022.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Neeraj Vashistha and Arkaitz Zubiaga. 2020. Online multilingual hate speech detection: Experimenting with hindi and english social media. *Information*, 12(1):5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Munsif Vengattil and Elizabeth Culliford. 2022. Facebook allows war posts urging violence against russian invaders. In *Reuters*.

Lígia Iunes Venturott and Patrick Marques Ciarelli. 2020. Data augmentation for improving hate speech detection on social networks. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '20, page 249–252, New York, NY, USA. Association for Computing Machinery.

Ligia Iunes Venturott and Patrick Marques Ciarelli. 2021. Application of Data Augmentation Techniques for Hate Speech Detection with Deep Learning. In G. Marreiros, F.S. Melo, N. Lau, H. Lopes Cardoso, and L.P. Reis, editors, *Progress in Artificial Intelligence. EPIA 2021. Lecture Notes in Computer Science*, volume 12981. Springer.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. 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, San Diego, California. Association for Computational Linguistics.

Yin Wenjie and Zubiaga Arkaitz. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. URL: https://arxiv.org/pdf/2102.08886.pdf, access date 15.02.2022.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Wiegand and Melanie Siegel. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, Vienna, Austria.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Youtube. 2022. Youtube hate speech definition. URL: https://support.google.com/youtube/answer/2801939?hl=en, access date 23.02.2022.

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 American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Ziqi Zhang, David Robinson, and Jonathan A. Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *ESWC*.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving Hate Speech Detection with Deep Learning Ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Data Collection and Annotation Process

A.1 Data Collection

Our data, originally unlabelled, stems from three different sources:

1. comments posted under online newspaper articles collected by our NGO partner (NGO)

2. online newspaper comments directly donated by three Swiss national online newspaper outlets (ON1, ON2 and ON3)

3. tweets collected using the Twitter API of "politically interested users", i.e. accounts following at least five Swiss newspapers or politicians.

The first two sources donated their data to us, while we collected tweets using the Twitter API. Particularly important for our study is source number two given that it includes published, moderated and deleted comments from three different media outlets. Each of the three ONs attracts different kinds of audiences in terms of age, and educational and sociopolitical backgrounds which influence the types of conversations, writing styles and uses of language more generally. Readers of and commentators in ON1 and ON2 display more similar characteristics than do those in ON3, as was also reflected in our post-deployment cross-dataset experiment results detailed in Table 4. Using all three sources thus adds to the diversity of our dataset.

A.2 Annotators

Annotators for (Swiss) German were 16 Bachelor and Master students in political science aged 20 to 27 years, of whom 15 are native speakers of (Swiss) German; annotators for French examples were three students from various disciplines in the natural sciences, all native speakers of French and 20 to 25 years of age. Differences in mother tongues and directions of study introduce some diversity, but overall, our annotators form a homogeneous group.

A.3 Annotator Instructions

All annotators attended a workshop introducing them to the annotation scheme depicted in Figure 1. As can be seen, annotators were instructed first to decide whether or not a comment or tweet contains hate speech or toxic speech and to enter the
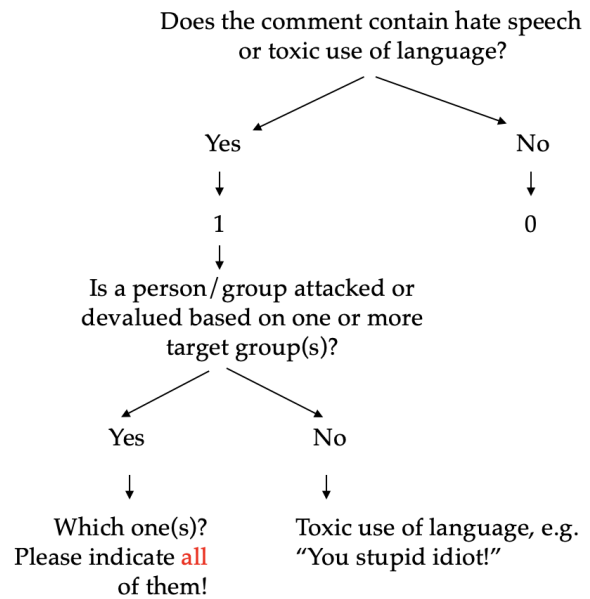
Figure 1: Annotation Scheme

corresponding number in the label column of their excel sheets. In case the comment or tweet contains hate speech, they were asked to indicate all groups against which the hate speech is targeted. Important to note in this scheme is that while a comment or tweet may be classified as targeting multiple groups, toxic speech is exclusive, meaning that once a comment or tweet is deemed to be toxic, it cannot be assigned to any other category. It is, however, still marked as 1 in the label column.

A.4 Iterative Labelling

The annotation process relied on iterative labelling as used by e.g. Kiela et al. (2021) and took full advantage of our human-in-the-loop hate speech classification pipeline with the exception that the human checks were performed by research assistants (RAs) and that retraining was carried out manually (cf. also Fanton et al., 2021).

The annotation process took place in five waves over a period of 12 months. Before the first wave, a set 0 of collected (Swiss) German comments was annotated by the NGO community. The first two waves of annotations each lasted 9 weeks. During the first, (Swiss) German data was annotated, during the second, French data. Since the initial annotations revealed that our data features many non-hate speech and borderline cases, but hardly any clear hate speech cases, samples distributed for annotation included only those comments and tweets which were classified as having the highest hate speech probability. By seeing increasingly

more instances of hate speech, the classifier gradually learnt to draw a clearer decision boundary between hate speech and non-hate speech.

At the beginning of the first two waves, a member of our team therefore annotated a sample of the classified comments with the highest hate speech probability. From then on, the two waves followed weekly cycles involving these steps:

1. (Re)training the model using a class-balanced set drawn from all the data annotated up to that point.

2. Making predictions on the (remaining) donated comments (first seven weeks) and tweets (final two weeks).

3. From the positively classified comments and tweets, a subset is drawn in descending order of hate speech probability and is divided into smaller sets.

4. The sets are distributed to RAs for annotation.

5. While the sets are being annotated, a team member carries out quality control on the sets annotated in the previous week.

6. The quality-controlled sets are added to the rest of the training set.

7. The process begins anew until the final set is annotated and checked.

For reasons of time and infrastructure, each comment and tweet was annotated by a single RA with the exception that in each weekly RA set, 60 identical comments or tweets were included for quality control. The third wave of annotations lasted for a week and was carried out on a set of (Swiss) German tweets. In this case, each tweet was annotated three times.

The fourth wave of annotations took place post-deployment as part of a bigger experiment on counter speech strategies on Twitter and lasted for 15 weeks. Given this setup, single annotations were carried out, with a member of our team carrying out second annotations, this time not just of the tweets with highest hate speech probability, but taking – for the first time – advantage of the full prediction probability scale, i.e. from 1.0 to 0.5. The results of this manner of proceeding are summarised in Table 6 and highlight that there is a direct correlation between the number of hate speech tweets,

their quality and the probability with which they were classified as hate speech: the further down the hate speech probability scale we go, the fewer actual hate speech tweets we find; and those we do find, are often contested, borderline or problematic cases.

The fifth wave of annotations also took place post-deployment, lasted for 9 weeks and involved three 3-week cycles through our human-in-the-loop hate speech classification pipeline. Each sample was annotated by two different RAs with a third RA making the final call on examples on which the first two could not agree.

In cases where multiple annotations were available for a single example, unweighted linear averaging was applied using majority ruling. To avoid falling into the doctrinal paradox, the resulting final annotations were manually checked for consistency during quality control, either by RAs themselves in the third and fourth waves or by a member of our team in the other waves.

The final result of this annotation process is, as shown in Table 12, a large corpus of over 422k unique data points annotated for the binary label hate speech, the 10 multilabel target groups and the binary category toxic speech. This corpus is the first of its kind and is, when compared to other benchmark hate speech and toxic language corpora (Waseem and Hovy, 2016; Davidson et al., 2017; Basile et al., 2019), one of – if not the – largest.

A.5 Quality Control

Quality control was carried out in terms of both quantity and quality. For the quantitative checks, we used Krippendorff's alpha to measure inter-rater reliability. Depending on the annotation wave, the results of this measurement oscillated between 0.3-0.4, which – while fairly low – is not surprising given the challenging and subjective nature of the task (Kwok and Wang, 2013; Ross et al., 2016; Waseem, 2016; Del Vigna et al., 2017; Malmasi and Zampieri, 2018; Mandl et al., 2019; Ousid-houm et al., 2019; Rani et al., 2020). Consequently, bi-weekly deliberation rounds were introduced to discuss difficult cases. In waves where this was not possible, peer respect was used, with a member of our team carrying out second annotations on the positively annotated comments and tweets. While this meant mostly checking that the annotation scheme was followed as closely as possible and that all annotators classified the same type of

hate speech in the same category, there were cases where a different annotation resulted after quality control. In addition to these quality checks, a subset of negatively classified comments with lowest non-hate speech probabilities was given out for annotation during the first wave (TN). The purpose of this set was to gain insight into the number of false negatives among the comments with lowest non-hate speech probability.

A.6 Challenges and Heterogeneity of the Swiss Hate Speech Corpus

Most challenges annotators experienced relate to cases which are so firmly embedded in the Swiss context that their negative connotations are only recognisable through subtexts. Examples of such expressions are listed in Table 7 and show that most of these instances target groups based on their nationality or political orientation.

Some of the contextual hate speech instances rely on the use of specific words (*rital*, *frouze*) or a distortion of words (*Balkonerin*) to designate a particular target group, while others imply a target group through expressing a certain type of behaviour or other distinct characteristic. The notions *(leased) BMW* and *Turnhosen*, for instance, are derogatory terms for people, mostly men, from the Balkans on the basis that many of them wear gym shorts and drive BMW cars; and the expression *grün-rote Wassermelone*, when used in a political context, attacks proponents of the green party by accusing them in a derogatory manner of pretending to be green, when in fact they are red, i.e. a socialist party.

In cases where RAs annotated comments, they were able to gauge the respective contexts more easily: comments sections are often not limited in the number of words or characters users can post, so individual comments are on average longer than tweets and provide more pieces of background information. RAs could also consult the title and text of the articles to gain additional context, which was not possible when annotating tweets. A further issue with tweets was that many of them were about events in Germany or Austria, using expressions and insults which are specific to these particular contexts and with which our RAs were not familiar.

While, with time and experience of these contexts, subtexts and jargon, RAs were able more easily to identify such difficult cases and the corresponding target groups, they also reported that

through continual exposure to hateful and toxic statements, they felt that they were becoming increasingly desensitised to them, an observation already described by Soral et al. (2018) and Barnidge et al. (2019). As a consequence, RAs noticed that there were cases which they would themselves annotate differently, once they re-read them at a later point in time.

RAs also reported that there remain many borderline cases, mostly involving implicit statements, terms which are potentially reclaimed by a minority group (Malmasi and Zampieri, 2018), strongly phrased personal opinions, sarcasm and irony, on which they could, even after discussing them, not agree. Another issue RAs faced concerns the diversity of languages and language varieties present in the dataset. Since hardly any guidelines are given on social media platforms as to the use of language, our dataset features a mix of the Swiss national languages, other languages commonly spoken in Switzerland (e.g. English and Portuguese) as well as the many dialects and varieties of Swiss German, whereby comments and tweets in German mostly use Standard German. Important to note is that both entire comments in a certain language or dialect variety and comments where several languages and dialect varieties are used simultaneously are present.

While we were able to filter tweets according to language and could minimise this phenomenon, this was not always possible for comments below online newspaper articles. This is why, in addition to the linguistic difficulties, like different vocabulary, sentence structures and so on, which this change in languages and varieties poses, there are occasionally (parts of) comments which are written in a different script corresponding to the language being used (e.g. Arabic, Cyrillic or Greek scripts), sometimes triggering a different direction of reading (cf. Arabic). In addition to these language specific changes, the dataset also displays many graphic alterations deliberately introduced by users who – presumably to game the system (Hosseini et al., 2017) – permute characters, words, sentence structures and make increased use of acronyms and abbreviations. This mix of codes and scripts was challenging for RAs who either decided not to annotate the comment or tweet in question or who spent a significant amount of time trying to research its content or translate it with the help of e.g. DeepL or Google Translate.

Lang	Hate Speech Example	Translation
G	Bin i dr einzig, wo gseht us welere Bevölkerigsgruppe dr chunt (*Turnhose*). Wiit muäs mr nied suächä, der *geleaste BMW* ist nicht weit.	Am I the only one who sees which people he belongs to (*gym shorts*). One does not have to look for long, the *leased BMW* can't be far.
	Diese *Balkonerin* soll sich in ihre heimat *Balkon* verschwindisieren.	That *female balcony* should return to her country *balcony*.
	grüne socken sind wie *wassermelonen*, aussen grün, innen rot.	*green socks* are like *watermelons*, green on the outside, red on the inside.
FR	Il faut aussi savoir si tu es *espingouin* ou *rital*?	You have to know whether you are *Spanish* or *Italian*?
	Elle aura ce qu'elle mérite et peut-être que ça lui apprendra à fermer sa grande g#*"© de *frouzes*!	She'll have what she deserves and maybe that'll teach her to shut her *French* mouth!
	de Genève les pleureuses *shadoks* sont de retour ?	from Geneva, the winy *French* are back?
	Le *seutch* tu peux le garder.	You can keep the *French*.
	Allez dire ça à ces foutus *bourbines*	Go tell this to those damn *Swiss Germans*

Table 7: Contextual Hate Speech

A.7 The Swiss Hate Speech Corpus in Numbers

The annotations generated in each wave of this process are summarised in Table 10. As can be seen, of the 442782 annotations, 112337 are hate or toxic speech (HS&Tox). Of these, 26815, i.e. approximately 24%, were deemed toxic speech (Tox) and 85522 are hate speech (HS) against a total of 109812 groups (All) from 1 or more of the 10 possible (multilabel) target groups, suggesting that each targeted comment or tweet is, on average, targeting approximately 1.3 groups. Differences in target group distribution emerge when comparing the ONs and Twitter as well as the two languages. The most targeted groups in the ONs are nationality followed by sex, politics and social status, and on Twitter politics followed by other targets, nationality and social status. Twitter too features significantly higher numbers of toxic speech than do the ONs.

Even though the groups which are most targeted are almost the same in both languages, the distribution of attacks among them varies, as shown in Table 13: from the total of targeted hate speech comments in the German ONs, approximately 39% feature hate speech against nationality, 32% against politics, 14% against other target groups, 13% against sex and 12% against social status compared to approximately 46% nationality, 28% sex, 20% politics and 17% social status in French. Similarly for Twitter: German has approximately 62% hate speech against politics, 18% against other targets and 14% against nationality and social status, while French has 49% against politics, 19% against nationality and 18% against sex and social status. These numbers suggest that French comments and tweets target politics to a lesser degree, and nationality, sex and social status more often than the Germans do. The same is true for impairments and appearance, while the reverse holds for the category 'other' where more German ON comments can be found. Despite the fact that the two languages feature similar numbers of attacks against age, gender and religion, these results confirm that language and sociocultural and political factors play a vital role in the frequency, distribution and nature of hate speech in Switzerland, and that varied distributions of them in the training set have the potential to bias the system.

B Results Tables

Description	Details	Studies
Input data	tweets	Waseem and Hovy (2016)
		Waseem (2016)
		Badjatiya et al. (2017)
		Davidson et al. (2017)
		Malmasi and Zampieri (2018)
		Pereira-Kohatsu et al. (2019)
	comments	Sood et al. (2012)
		Kwok and Wang (2013)
		Saleem et al. (2017)
		Del Vigna et al. (2017)
		Rani et al. (2020)
Preprocessing	automatic translation into English	Aluru et al. (2020)
Languages	monolingual (mostly English)	Malmasi and Zampieri (2018)
		Rani et al. (2020)
		Aljero and Dimililer (2021)
	monolingual non-English	Del Vigna et al. (2017)
		Pereira-Kohatsu et al. (2019)
		Çöltekin (2020)
		Alshaalan and Al-Khalifa (2020)
		Taradhita and Putra (2021)
	multilingual	Chiril et al. (2019)
		Ousidhoum et al. (2019)
		Aluru et al. (2020)
		Rani et al. (2020)
		Vashistha and Zubiaga (2020)
		Nozza et al. (2020)
		Tita and Zubiaga (2021)
Classifier types	binary	Malmasi and Zampieri (2018)
		Rani et al. (2020)
		Aljero and Dimililer (2021)
	multilabel/-class	Waseem and Hovy (2016)
		Ross et al. (2016)
		Davidson et al. (2017)
		Salminen et al. (2018)
Model architectures	statistical	Saleem et al. (2017)
		Pereira-Kohatsu et al. (2019)
	deep learning and transformers	Badjatiya et al. (2017)
		Zhang et al. (2018)
		Zampieri et al. (2019)
		Mozafari et al. (2019)
		Geet D'Sa et al. (2020)
		Kupi et al. (2021)
		Rani et al. (2020)
		Malik et al. (2022)
	single and ensemble models	Malmasi and Zampieri (2018)
		Kokatnoor and Krishnan (2020)
		Gao and Huang (2017)
		Zimmerman et al. (2018)
		Melton et al. (2020)
		Aljero and Dimililer (2021)
	sentiment analysis	Thelwall et al. (2010)
		Sood et al. (2012)
		Van Hee et al. (2015)
Feature engineering	surface-level features	Mondal et al. (2017)
	word embeddings	Djuric et al. (2015)
		Nobata et al. (2016)
	LDA	Saleem et al. (2017)
	additional features to text	Davidson et al. (2017)
		Gao and Huang (2017)
		Chaudhry and Lease (2020)
Categories	hate speech, toxic, profanity	Davidson et al. (2017)
		Malmasi and Zampieri (2018)
	specific target groups	Kwok and Wang (2013)
		Burnap and Leighton Williams (2014)
		Burnap and Williams (2016)
		Waseem and Hovy (2016)
		Fersini et al. (2018)
		Pamungkas et al. (2018)
		Jaki and De Smedt (2019)
		Chiril et al. (2019)
		Basile et al. (2019)
		Duzha et al. (2021)
	voice-for-the-voiceless	Palakodety et al. (2020)

Table 8: Related Work

Competition	Evaluation Metric	Binary	Binary Classes	Multiclass	Multi Classes	Dataset	Language
SemEval (Basile et al., 2019)	macro-averaged F1	35.0-65.1	(non-)hateful	15.9-57.0	individual, generic, (non-)aggressive	Twitter	English
SemEval (Basile et al., 2019)	macro-averaged F1	49.3-73.0	(non-)hateful	42.8-70.5	individual, generic, (non-)aggressive	Twitter	Spanish
GermEval2018 (Wiegand and Siegel, 2018)	macro F1	49.0-76.8	offense, other	32.1-52.7	profanity, insult, abuse, other	Twitter	German
GermEval2019 (Struss et al., 2019)	macro F1	54.9-76.9	offense, other	36.8-53.6	profanity, insult, abuse, other	Twitter	German
GermEval2019 (Struss et al., 2019)	macro F1	55.4-73.1	implicit, explicit			Twitter	German
GermEval2021 (Risch et al., 2021)	macro F1	36.0-71.8	toxic comments			Facebook	German
GermEval2021 (Risch et al., 2021)	macro F1	61.4-70.0	engaging comments			Facebook	German
GermEval2021 (Risch et al., 2021)	macro F1	59.7-76.3	fact-claiming comments			Facebook	German
Evalita2018 (Fersini et al., 2018)	accuracy	76.5-84.4	(non-)misogynistic	29.2-50.1	stereotype, objectification, dominance, derailing, sexual harassment, discredit, individual, generic	Twitter	Italian
Evalita2018 (Fersini et al., 2018)	accuracy	50.0-70.4	(non-)misogynistic	23.2-40.6	stereotype, objectification, dominance, derailing, sexual harassment, discredit, individual, generic	Twitter	English
HASOC (Mandl et al., 2019) (Mandl et al., 2020, 2021)	macro F1	50.1-83.1	(non-)hateful	45.0-54.5	hate speech, offensive, profane	Twitter	English
	macro F1	45.8-51.1	(un)targeted			Twitter	English
HASOC (Mandl et al., 2019) (Mandl et al., 2020, 2021)	macro F1	46.0-81.5	(non-)hateful	06.2-58.1	hate speech, offensive, profane	Twitter	Hindi
	macro F1	49.7-57.5	(un)targeted			Twitter	Hindi
HASOC (Mandl et al., 2019, 2020)	macro F1	47.9-61.6	(non-)hateful	24.0-34.7	hate speech, offensive, profane	Twitter	German
HASOC (Mandl et al., 2021)	macro F1	53.9-91.4	(non-)hateful		hate speech, offensive, profane	Twitter	Marathi

Table 9: Summary of Competition Results

Wave	Week	Lang	Data	Examples			Targets												Tox	HS
				Total	nonHS	HS&Tox	Sex	Age	Gen.	Rel.	Nat.	Imp.	Stat.	Pol.	App.	Other	All		
0	NGO	G	NGO	12860	12494	366	6	0	0	0	12	0	10	0	0	0	28	338	28
1	0	G	ON1	25665	24518	1147	230	28	34	19	448	0	66	351	0	0	1176	92	1055
	1	G	ON1	22800	21866	934	192	52	70	52	487	14	49	88	18	98	1120	105	829
	2	G	ON1	16800	16019	781	122	41	41	34	282	22	60	103	42	33	780	129	652
	3	G	ON1	22800	21244	1556	159	75	40	34	567	64	140	211	138	50	1478	325	1231
	4	G	ON1	23800	22087	1713	140	45	49	55	600	16	146	305	134	158	1648	368	1345
	5	G	ON1	23800	22547	1253	51	61	14	6	303	15	90	82	50	164	836	515	738
	6	G	ON1	26000	22927	3073	216	133	40	89	1285	21	330	443	109	125	2791	724	2349
	7	G	ON1	26000	22007	3993	344	154	56	80	1288	34	477	665	155	164	3417	1117	2876
	TN	G	ON1	15000	14872	128	0	0	0	0	8	0	23	22	0	31	84	55	73
	8	G	Tweets	29798	25286	4512	218	89	25	147	621	79	359	1416	78	77	3109	1898	2614
	9	G	Tweets	5000	4487	513	27	10	6	15	74	6	33	181	5	8	365	196	317
2	0	FR	ON1	4000	2780	1220	115	29	28	33	216	8	59	102	18	1	609	767	453
	1	FR	ON1	16178	7718	8460	2087	270	289	221	2945	222	693	1107	502	106	8442	2094	6366
	2	FR	ON1	16353	10066	6287	1747	217	188	161	2210	213	626	1092	383	50	6887	929	5358
	3	FR	ON1	13000	9391	3609	676	112	104	54	1059	108	364	596	191	26	3290	1027	2582
	4	FR	ON1	12001	4244	7757	1566	106	141	144	4894	150	1319	1229	370	25	9944	390	7367
	5	FR	ON1	11447	5375	6072	1937	88	130	145	1969	130	1103	848	300	71	6721	802	5270
	6	FR	ON1	9051	4045	5006	1262	74	101	120	1490	113	1188	1057	174	49	5628	608	4398
	7	FR	ON1	10593	6827	3766	608	51	36	62	810	80	850	958	173	91	3719	829	2937
	8	FR	Tweets	9993	8417	1576	241	19	33	116	239	55	284	469	64	39	1559	519	1057
	9	FR	Tweets	16997	14093	2904	254	35	52	145	275	85	220	844	41	173	2124	1263	1641
3	1	G	Tweets	15411	6854	8557	666	172	143	468	1210	104	380	3108	112	605	6968	3095	5462
4	1-15	G	Tweets	33822	15395	18427	718	192	149	222	1393	86	2411	9449	136	3527	18283	3937	14490
5	1-3	G	ON1	11304	3370	7934	771	110	143	97	1880	38	528	2116	170	947	6800	2545	5389
	4-6	G	ON2	6267	1244	5023	440	78	119	128	875	5	397	2383	56	1117	5598	1003	4020
	7-9	G	ON3	6042	272	5770	495	102	169	89	956	15	493	2418	100	1571	6408	1145	4625
Total				442782	330445	112337	15288	2343	2200	2736	28396	1683	12698	31643	3519	9306	109812	26815	85522

Table 10: Results Annotation Process

			Base			Tuned		
Lang	**Dataset**	**Model**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
G	Comments	GELECTRA$_{Base}$	44.4	47.4	39.3	76.0	75.5	75.4
G	Comments	GBERT$_{Base}$	44.1	49.8	33.7	75.7	75.4	75.4
G	Comments	mBERT$_{Base}$	25.0	50.0	33.3	74.7	74.4	74.4
G	Comments	XLM-RoBERTa$_{Base}$	25.0	50.0	33.3	75.7	75.0	74.9
G	Comments	Twitter-XLM-RoBERTa$_{Base}$	41.7	50.0	33.3	75.4	75.1	75.0
FR	Comments	flauBERT$_{Base}$	49.6	49.6	49.3	66.5	66.5	66.4
FR	Comments	french RoBERTa$_{Base}$	46.7	49.1	37.9	62.3	62.3	62.3
FR	Comments	mBERT$_{Base}$	58.0	50.0	33.4	67.4	67.2	66.9
FR	Comments	XLM-RoBERTa$_{Base}$	49.5	49.9	37.3	67.2	67.2	67.1
FR	Comments	Twitter-XLM-RoBERTa$_{Base}$	37.4	50.0	33.3	67.2	67.1	67.0
G & FR	Comments	mBERT$_{Base}$	40.3	49.9	33.4	79.8	79.6	79.6
G & FR	Comments	XLM-RoBERTa$_{Base}$	50.5	50.5	50.0	79.9	79.9	79.8
G & FR	Comments	Twitter-XLM-RoBERTa$_{Base}$	50.5	50.5	50.0	79.9	79.9	79.8
G	Tweets	GELECTRA$_{Base}$	52.7	51.0	42.1	79.5	79.3	79.3
G	Tweets	GBERT$_{Base}$	56.0	55.6	54.9	78.8	78.8	78.8
G	Tweets	mBERT$_{Base}$	25.0	50.0	33.3	78.2	78.2	78.2
G	Tweets	XLM-RoBERTa$_{Base}$	61.4	50.0	33.4	78.6	78.3	78.3
G	Tweets	Twitter-XLM-RoBERTa$_{Base}$	50.0	50.0	33.3	78.7	78.7	78.7
FR	Tweets	flauBERT$_{Base}$	51.4	51.4	51.3	OF	OF	OF
FR	Tweets	french RoBERTa$_{Base}$	58.4	50.2	34.0	OF	OF	OF
FR	Tweets	mBERT$_{Base}$	55.0	50.0	33.5	OF	OF	OF
FR	Tweets	XLM-RoBERTa$_{Base}$	25.0	50.0	33.3	OF	OF	OF
FR	Tweets	Twitter-XLM-RoBERTa$_{Base}$	25.0	50.0	33.3	OF	OF	OF
G & FR	Tweets	mBERT$_{Base}$	52.7	51.7	46.8	76.4	76.4	76.4
G & FR	Tweets	XLM-RoBERTa$_{Base}$	25.0	50.0	33.3	78.1	77.9	77.9
G & FR	Tweets	Twitter-XLM-RoBERTa$_{Base}$	25.0	50.0	33.3	77.8	77.4	77.4
G	Comments & Tweets	GELECTRA$_{Base}$	54.3	52.7	47.7	76.5	76.4	76.4
G	Comments & Tweets	GBERT$_{Base}$	46.5	47.7	42.8	76.6	76.5	76.5
G	Comments & Tweets	mBERT$_{Base}$	50.5	50.3	44.1	74.0	73.9	73.9
G	Comments & Tweets	XLM-RoBERTa$_{Base}$	52.5	51.3	44.7	75.2	75.2	75.2
G	Comments & Tweets	Twitter-XLM-RoBERTa$_{Base}$	57.9	50.0	33.6	75.4	75.3	75.2
FR	Comments & Tweets	flauBERT$_{Base}$	48.2	48.4	46.5	67.6	67.1	66.9
FR	Comments & Tweets	french RoBERTa$_{Base}$	25.0	50.0	33.3	62.6	62.6	62.6
FR	Comments & Tweets	mBERT$_{Base}$	51.6	50.0	33.8	67.3	67.3	67.3
FR	Comments & Tweets	XLM-RoBERTa$_{Base}$	50.0	50.0	33.3	67.4	67.3	67.3
FR	Comments & Tweets	Twitter-XLM-RoBERTa$_{Base}$	49.0	49.8	36.8	68.2	68.0	68.0
G & FR	Comments & Tweets	mBERT$_{Base}$	25.0	50.0	33.3	79.3	79.2	79.2
G & FR	Comments & Tweets	XLM-RoBERTa$_{Base}$	25.0	50.0	33.3	71.2	71.1	71.1
G & FR	Comments & Tweets	Twitter-XLM-RoBERTa$_{Base}$	25.0	50.0	33.3	71.2	71.1	71.1

Table 11: BERT-Models Final Results

Lang	Data	Examples			Targets											Tox	HS
		Total	nonHS	HS&Tox	Sex	Age	Gender	Rel.	Nat.	Imp.	Status	Pol.	App.	Other	All		
G	Comments	225232	201445	23787	2288	751	574	535	7032	206	2098	5846	813	2609	22752	5629	18158
G	Tweets	83851	51885	31966	1628	463	324	848	3295	275	3179	14145	331	4213	28701	9107	22859
FR	Comments	86213	47307	38906	9062	890	976	862	14737	955	5492	6424	1961	389	41748	6910	31996
FR	Tweets	26750	22333	4417	492	54	83	257	511	140	493	1302	105	212	3649	1749	2668
Total	Both	422046	322970	99076	13470	2158	1957	2502	25575	1576	11262	27717	3210	7423	96850	23395	75681

Table 12: Unique Annotations in the Swiss Hate Speech Corpus

	ONs				Twitter			
	G HS = 18 158		FR HS = 31 996		G HS = 22 859		FR HS = 2668	
Target	HSGroup	%	HSGroup	%	HSGroup	%	HSGroup	%
Sex	2288	13	9062	28	1628	7	492	18
Age	751	4	890	3	463	2	54	2
Gender	574	3	976	3	324	1	83	3
Rel.	535	3	862	3	848	4	257	10
Nat.	7032	39	14737	46	3295	14	511	19
Imp.	206	1	955	3	275	1	140	5
Status	2098	12	5492	17	3179	14	493	18
Pol.	5846	32	6424	20	14145	62	1302	49
App.	813	4	1961	6	331	1	105	4
Other	2609	14	389	1	4213	18	212	8

Table 13: Percentages of Hate Speech per Multilabel Target Group