# Robust Hate Speech Detection in Social Media:
# A Cross-Dataset Empirical Evaluation

**Dimosthenis Antypas** and **Jose Camacho-Collados**
Cardiff NLP, School of Computer Science and Informatics
Cardiff University, United Kingdom
{AntypasD,CamachoColladosJ}@cardiff.ac.uk

## Abstract

The automatic detection of hate speech online is an active research area in NLP. Most of the studies to date are based on social media datasets that contribute to the creation of hate speech detection models trained on them. However, data creation processes contain their own biases, and models inherently learn from these dataset-specific biases. In this paper, we perform a large-scale cross-dataset comparison where we fine-tune language models on different hate speech detection datasets. This analysis shows how some datasets are more generalizable than others when used as training data. Crucially, our experiments show how combining hate speech detection datasets can contribute to the development of robust hate speech detection models. This robustness holds even when controlling by data size and compared with the best individual datasets.

## 1 Introduction

Social media has led to a new form of communication that has changed how people interact across the world. With the emergence of this medium, hateful conduct has also found a place to propagate online. From more obscure online communities such as 4chan (Knuttila, 2011) and Telegram rooms (Walther and McCoy, 2021) to mainstream social media platforms such as Facebook (Del Vigna12 et al., 2017) and Twitter (Udanor and Anyanwu, 2019), the spread of hate speech is an on going issue.

Hate speech detection is a complex problem that has received a lot of attention from the Natural Language Processing (NLP) community. It shares a lot of challenges with other social media problems (emotion detection, offensive language detection, etc), such as an increasingly amount of user generated content, unstructured (Elsayed et al., 2019) and constantly evolving text (Ebadi et al., 2021), and the need of efficient large scale solutions. When dealing with hate speech in particular,

one has to consider the sensitivity of the topics, their wide range (e.g. sexism, sexual orientation, racism), and their evolution through time and location (Matamoros-Fernández and Farkas, 2021). Understanding the extent of the problem and tracking hate speech online through automatic techniques can therefore be part of the solution of this ongoing challenge. One way to contribute to this goal is to both improve the current hate speech detection models and, crucially, the data used to train them.

The contributions of this paper are twofold. First, we provide a summary and unify existing hate speech detection datasets from social media, in particular Twitter. Second, we analyse the performance of language models trained on all datasets, and highlight deficiencies in generalisation across datasets, including the evaluation in a new independently-constructed dataset. Finally, as a practical added value stemming from this paper, we share all the best models trained on the unification of all datasets, providing a relatively small-size hate speech detection model that is generalisable across datasets.[1]

**Content Warning** The article contains examples of hateful and abusive language. The first vowel in hateful slurs, vulgar words, and in general profanity language is replaced with an asterisk (*).

## 2 Related Work

Identifying hate speech in social media is an increasingly important research topic in NLP. It is often framed as a classification task (binary or multiclass) and through the years various machine learning and information sources approaches have been

---

[1]The best binary hate speech detection model is available at https://huggingface.co/cardiffnlp/twitter-roberta-base-hate-latest; the multiclass hate speech detection model identifying target groups is available at https://huggingface.co/cardiffnlp/twitter-roberta-base-hate-multiclass-latest. These models have been integrated into the TweetNLP library (Camacho-Collados et al., 2022).

utilised (Mullah and Zainon, 2021; Ali et al., 2022; Khanday et al., 2022; del Valle-Cano et al., 2023). A common issue of supervised approaches lies not necessarily with their architecture, but with the existing hate speech datasets that are available to train supervised models. It is often the case that the datasets are focused on specific target groups (Grimminger and Klinger, 2021), constructed using some specific keyword search terms (Waseem and Hovy, 2016; Zampieri et al., 2019), or have particular class distributions (Basile et al., 2019) that leads to a training process that may or may not generalise. For instance, Florio et al. (2020) analysed the temporal aspect of hate speech, and demonstrate how brittle hate speech models are when evaluated on different periods. Recent work has also shown that there is a need to both focus on the resources available and also try to expand them in order to develop robust hate speech classifiers that can be applied in various context and in different time periods (Bourgeade et al., 2023; Bose et al., 2022).

In this paper, we perform a large-scale evaluation to analyse how generalisable supervised models are depending on the underlying training set. Then, we propose to mitigate the relative lack of generalisation by using datasets from various sources and time periods aiming to offer a more robust solution.

## 3 Data

In this section, we describe the data used in our experiments. First, we describe existing hate speech datasets in Section 3.1. Then, we unify those datasets and provide statistics of the final data in Section 3.2

### 3.1 Hate Speech datasets

In total, we collected 13 datasets related to hate speech in social media. The datasets selected are diverse both in content, different kind of hate speech, and in a temporal aspect.

**Measuring hate speech (MHS)** *MHS* (Kennedy et al., 2020; Sachdeva et al., 2022) consists of 39,565 social media (YouTube, Reddit, Twitter) manually annotated comments. The coders were asked to annotate each entry on 10 different attributes such as the presence of sentiment, respect, insults and others; and also indicate the target of the comment (e.g. age, disability). They use Rasch measurment theory (Rasch, 1960) to aggregate the annotators' rating in a continuous value that indicates the hate score of the comment.

**Call me sexist, but (CMS)** This dataset of 6,325 entries (Samory et al., 2021) focuses on the aspect of sexism and includes social psychology scales and tweets extracted by utilising the "Call me sexist, but" phrase. The authors also include two other sexism datasets (Jha and Mamidi, 2017; Waseem and Hovy, 2016) which they re-annotate. Each entry is annotated by five coders and is labelled based on its content (e.g. sexist, maybe-sexist) and phrasing (e.g. civil, uncivil).

**Hate Towards the Political Opponent (HTPO)** *HTPO* (Grimminger and Klinger, 2021) is a collection of 3,000 tweets related to the 2020 USA presidential election. The tweets were extracted using a set of keywords linked to the presidential and vice presidential candidates and each tweet is annotated for stance detection (in favor of/against the candidate) and whether it contains hateful language or not.

**HateX** *HateX* (Mathew et al., 2021) is a collection of 20,148 posts from Twitter and Gab extracted by utilising relevant hate lexicons. For each entry, three annotators are asked to indicate: (1) the existence of hate speech, offensive speech, or neither of them, (2) the target group of the post (e.g. Arab, Homosexual), and (3) the reasons for the label assigned.

**Offense** The *Offense* dataset (Zampieri et al., 2019) contains 14,100 tweets extracted by utilising a set of keywords and categorises them in three levels: (1) offensive and non-offensive; (2) targeted/untargeted insult; (3) targeted to individual, group, or other.

**Automated Hate Speech Detection (AHSD)** In this dataset, (Davidson et al., 2017) the authors utilise a set of keywords to extract 24,783 tweets which are manually labelled as either hate speech, offensive but not hate speech, or neither offensive nor hate speech.

**Hateful Symbols or Hateful People? (HSHP)** This is a collection (Waseem and Hovy, 2016) of 16,000 tweets extracted based on keywords related to sexism and racism. The tweets are annotated as on whether they contain racism, sexism or neither

of them by three different annotators.[2]

**Are You a Racist or Am I Seeing Things? (AYR)**
This dataset (Waseem, 2016) is an extension of *Hateful Symbols or Hateful People?* and adds the "both" (sexism and racism) as a potential label. Overlapping tweets were not considered.

**Multilingual and Multi-Aspect Hate Speech Analysis (MMHS)** MMHS (Ousidhoum et al., 2019) contains hateful tweets in three different languages (English, French, Arabic). Each tweet has been labelled by three annotators on five different levels: (1) directness, (2) hostility (e.g. abusive, hateful), (3) target (e.g. origin, gender), (4) group (e.g. women, individual) and (5) annotator emotion (disgust, shock, etc). A total of 5,647 tweets are included in the dataset.

**HatE** *HatE* (Basile et al., 2019) consists of English and Spanish tweets (19,600 in total) that are labelled on whether they contain hate speech or not. The tweets in this dataset focus on hate speech towards two groups: (1) immigrants and (2) women.

**HASOC** This dataset (Mandl et al., 2019) contains 17,657 tweets in Hindi, German and English which are annotated on three levels: (1) whether they contain hate-offensive content or not; (2) in the case of hate-offensive tweets, whether a post contains hate, offensive, or profane content/words; (3) on the nature of the insult (targeted or un-targeted).

**Detecting East Asian Prejudice on Social Media (DEAP)** This is a collection of 20,000 tweets (Vidgen et al., 2020) focused on East Asian prejudice, e.g. Sinophobia, in relation to the COVID-19 pandemic. The annotators were asked to labelled each entry based on five different categories (hostility, criticism, counter speech, discussion, non-related) and also indicate the target of the entry (e.g. Hong Kongers, China).

**Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior (LSC)** The dataset (Founta et al., 2018) consists of 80,000 tweets extracted using a boosted random sample technique. Each tweet is labelled as either offensive, abusive, hateful, aggressive, cyberbullying or normal.

---

[2]A subset of the dataset is included in the *Call me sexist, but* and is not considered.

## 3.2 Unification

Even though all of the datasets that were collected revolve around hate speech, there are major differences among them in terms of both format and content. We attempt to unify the datasets by standarizing their format and combining the available content into two settings: (1) binary hate speech classification and (2) a multiclass classification task including the target group. We note that in cases where the original annotation results were provided, we decided to assign a label if at least two of the coders agree on it and not necessarily the majority of them. This approach can lead to a more realistic dataset and contribute in creating more robust systems (Antypas et al., 2022; Mohammad et al., 2018).

### 3.2.1 Initial preprocessing

For each dataset collected, a simple preprocessing pipeline is applied. Firstly, any non-Twitter content is removed; despite the similarities between the content shared in various social media (e.g. internet slang, emojis), Twitter displays unique characteristics, such as the concept of retweets and shorter texts, which differentiate it from other platforms such as Reddit or Youtube (Smith et al., 2012). Moreover, as our main consideration is hate speech in the English language, we exclude any non-English subset of tweets, and also verify the language by using a fastText based language identifier (Bojanowski et al., 2017). Finally, considering that some datasets in this study utilise similar keywords to extract tweets, we remove near duplicated entries to avoid any overlap between them. This is accomplished by applying a normalisation step where entries that are considered duplicated based on their lemmmatised form are ignored. Also, all URLs and mentions are removed.

As a final note, three of the datasets (*HSHP, AYR, LSC*) were dehydrated using the Twitter API since only their tweet IDs and their labels were publicly available. Unfortunately, a significant number of tweets ($\approx 10,000$) were no longer available from the API.

### 3.2.2 Binary Setting

The majority of the datasets collected are either set as a binary hate classification task and no further preprocessing is applied (*HTPO*), or offer a more fine-grained classification of hate speech (e.g. *HateX, CMS*) where we consider all "hate" subclasses as one. In general, when a dataset focuses on a spe-

cific type of hate speech (e.g. sexism) we map it as hate speech. Notable exceptions are: (1) The *MSH* dataset, where a continues hate score is provided which is transformed into a binary class according to the mapping proposed by the original authors. (2) Datasets that consist of offensive speech but also provide information about the target of the tweet. In these cases, (*Offense*), we consider only entries that are classified as offensive and are targeting a group of people and not individuals. Our assumption is that offensive language towards a group of people is highly likely to target protected characteristics and thus be classified as hate speech. (3) Finally, only entries classified as hate speech were considered in datasets where there is a clear distinction between hate, offensive, or profound speech (*LSC*, *AHSD*, *HASOC*). All data labelled as normal or not-hateful are also included as *not-hate* speech.

### 3.2.3 Multiclass Setting

Having established our binary setting, we aggregated the available datasets aiming to construct a more detailed hate speech classification task. As an initial step, all available hate speech sub-classes present were considered. However, this led to a very detailed but sparse hate taxonomy, with 44 different hate speech categories, but with only a few entries for some of the classes (e.g. "economic" category with only four tweets present). Aiming to create an easy-to-use and extendable data resource, several categories were grouped together. All classes related to ethnicity (e.g. Arab, Hispanic) or immigration were grouped under *racism*, while religious categories (e.g. Muslim, Christian) were considered separately. Categories related to sexuality and sexual orientation (e.g. heterosexual, homosexual) were also grouped in one class, and tweets with topics regarding gender (men, women) constitute the *sexism* class. Finally, all entries labelled as "not-hate" speech were also included. To keep our dataset relatively balanced we also ignored classes that constitute less than 1% of the total hate speech data. Overall, the multiclass setting proposed consists of 7 classes: *Racism, Sexism, Disability, Sexual orientation, Religion, Other*, and *Not-Hate*. It is worth noting that tweets falling under the *Other* class do not belong to any of the other five hate speech classes.

### 3.2.4 Statistics and Data Splits

In total, we collected 83,230 tweets, from 13 different datasets (Table 1), of which only 33% are classified as hate speech. This unified dataset may seem imbalanced but it is commonly assumed that only around 1% of the content shared on social media contains hate speech (Pereira-Kohatsu et al., 2019). When considering the multiclass setting, the hate speech percentage decreases even more with only 26% of tweets labelled as a form of hate speech, with the *religion* class being the least popular with only 709 entries.

The data in both settings (binary & multiclass) are divided into train and test sets using a stratified split to ensure class balance between the splits (Table 2). In general, for each dataset present, we allocate 70% as training data, 10% as validation, and 20% as test data. Exceptions to the aforementioned approach are datasets where the authors provide a preexisting data split which we use.

## 4 Evaluation

We present our main experimental results comparing various language models trained on single datasets and in the unified dataset presented in the previous section.

### 4.1 Experimental Setting

**Models.** For our experiments we rely on four language models of a similar size, two of them being general-purposes and the other two specialized on social media: BERT-base (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019) as general-purpose models; and BERTweet (Nguyen et al., 2020) and TimeLMs-21 (Loureiro et al., 2022) as language models specialized on social media, and particularly Twitter. There is an important difference between BERTweet and TimeLMs-21: since BERTweet was trained from scratch, TimeLMs-21 used the RoBERTa-base checkpoint as initialization and then continued training on a Twitter corpus. An SVM classifier is also utilized as a baseline model.

**Settings.** Aiming to investigate the effect of a larger and more diverse hate speech training corpus on various types of hate speech, we perform an evaluation on both the binary and multiclass settings described in Section 3.2. Specifically, for the binary setting we fine-tune the models selected first on each individual dataset, and secondly while using the unified dataset created. For the multiclass

| Dataset | Binary | | Multiclass | | | | | |
|---|---|---|---|---|---|---|---|---|
| | hate | not-hate | racism | sexism | sexual orientation | disability | religion | other |
| HatE | 5303 | 7364 | 2474 | 2829 | - | | | |
| MHS | 2485 | 5074 | 735 | 784 | 251 | 21 | 246 | 10 |
| DEAP | 3727 | 105 | 3727 | - | | | | |
| CMS | 1237 | 10861 | - | 1237 | - | | | |
| Offense | 1142 | 12547 | - | | | | | |
| HateX | 2562 | 5678 | 757 | 492 | 407 | 30 | 239 | 143 |
| LSC | 889 | 1267 | - | | | | | |
| MMHS | 5392 | - | 472 | 764 | 512 | 1387 | 224 | 2033 |
| HASOC | 1237 | 4348 | - | | | | | |
| AYR | 393 | 1246 | 42 | 343 | - | | | |
| AHSD | 1363 | 4088 | - | | | | | |
| HTPO | 351 | 2647 | - | | | | | |
| HSHP | 1498 | 426 | 9 | 1489 | - | | | |
| **Total** | **27,579** | **55,651** | **8,216** | **7,938** | **1,170** | **1,438** | **709** | **2,186** |

Table 1: Distribution of tweets gathered across hate speech datasets, including those where the target information is available (multiclass).

| Dataset | train | | test | |
|---|---|---|---|---|
| | not-hate | hate | not-hate | hate |
| AHSD | 3270 | 1090 | 818 | 273 |
| AYR | 996 | 314 | 250 | 79 |
| CMS | 8688 | 989 | 2173 | 248 |
| DEAP | 84 | 2981 | 21 | 746 |
| HASOC* | 3489 | 1113 | 859 | 124 |
| HSHP | 341 | 1197 | 85 | 301 |
| HTPO* | 2106 | 292 | 541 | 59 |
| HatE* | 5757 | 4197 | 1607 | 1106 |
| HateX | 4542 | 2050 | 1136 | 512 |
| LSC | 1013 | 711 | 254 | 178 |
| MHS | 4058 | 1988 | 1016 | 497 |
| Offense* | 10037 | 913 | 2510 | 229 |
| **All** | **44,381** | **17,835** | **11,270** | **4,352** |

Table 2: Binary class distribution in train and test splits of the unified hate speech datasets. * indicates datasets where preexisting train/test splits were available and retrieved.

setting, we considered the unified and the HateX dataset, which includes data for all classes. In total, we fine-tuned 54 different binary[3] and 8 multiclass models.

**Training.** The implementations provided by Hugging Face (Wolf et al., 2020) are used to train and evaluate all language models, while we utilise Ray Tune (Liaw et al., 2018) along with HyperOpt (Bergstra et al., 2022) and Adaptive Successive Halving (Li et al., 2018) for optimizing the learning rate, warmup steps, number of epochs, and batch size, hyper-parameteres of each model.[4]

**Evaluation metrics.** The macro-averaged F1 score is reported and used to compare the performance of the different models. Macro-F1 is commonly used in similar tasks (Basile et al., 2019; Zampieri et al., 2020) as it provides a more concrete view on the performance of each model.

### 4.2 Datasets

For training and evaluation, we use the splits described in Section 3.2.4. As described above, for each language model we trained on each dataset training set independently, and in the combination of all dataset-specific training sets. The results on the combination of all datasets are averaged across each dataset-specific test set (AVG), i.e., each dataset is given the same weight irrespective of its size. In addition to the datasets presented in Section 3.1, we constructed an independent test set (Indep) to test the robustness of models outside existing datasets.

**Independent test set (Indep).** This dataset was built by utilising a set of keywords related to the *International Women's Day* and *International Day Against Homophobia, Transphobia and Biphobia* and extracting tweets from the respected days of 2022. Then, these tweets were manually annotated

---

[3]MMHS dataset was used only for the training/evaluation of the unified dataset as it is lacking the *not-hate* class

[4]Optimal hyperparameters can be found in Table 5 in the Appendix

| Model | Train Data | Size | HatE | MHS | DEAP | CMS | Off. | HateX | LSC | HASOC | AYR | AHSD | HTPO | HSHP | AVG | Indep |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BERTweet | All | 58213 | 57.1 | 87.7 | **57.7** | **82.4** | 59.4 | **75.1** | 61.5 | 59.4 | 85.5 | **90.2** | 59.5 | **65.4** | 70.1 | 61.0 |
| | All* | 5290 | 51.1 | 80.5 | 53.7 | 73.1 | **60.8** | 67.3 | 72.1 | **63.9** | 85.6 | 85.4 | 67.6 | 62.1 | 68.6 | **69.2** |
| | MHS | 5291 | **65.5** | **89.3** | 13.3 | 50.6 | 53.2 | 69.6 | 58.8 | 58.0 | 66.8 | 78.8 | **67.7** | 28.6 | 58.3 | 58.6 |
| TimeLMs | All | 58213 | 54.2 | **86.6** | 68.0 | 79.7 | 56.9 | 74.8 | 59.1 | 63.2 | 87.2 | **89.4** | 65.2 | **64.5** | 70.7 | 63.7 |
| | All* | 1146 | 48.3 | 74.9 | 49.3 | 69.3 | 54.7 | 59.7 | **63.8** | 63.8 | 82.3 | 79.9 | 59.6 | 63.0 | 64.0 | **70.6** |
| | AYR | 1147 | **61.0** | 71.4 | 9.8 | 63.5 | 52.5 | 56.3 | 60.9 | 63.6 | **87.7** | 80.7 | **66.8** | 57.9 | 61.0 | 59.3 |
| RoBERTa | All | 58213 | 52.3 | **85.9** | 66.6 | 79.9 | 54.7 | 73.8 | 59.5 | 60.8 | 87.0 | **89.8** | 64.4 | 61.4 | **69.7** | 56.2 |
| | All* | 1146 | **56.0** | 73.7 | 53.2 | 64.2 | 53.0 | 48.9 | 70.2 | 65.8 | 74.3 | 74.1 | 58.9 | 61.0 | 62.8 | **78.3** |
| | AYR | 1147 | 54.8 | 63.8 | 17.5 | 69.8 | 55.2 | 50.1 | 57.7 | 63.4 | 86.3 | 81.9 | **64.6** | 55.6 | 60.1 | 53.8 |
| BERT | All | 58213 | 52.3 | **84.0** | 49.3 | 79.7 | 56.8 | 74.1 | 56.9 | **60.9** | 85.2 | 89.6 | 60.5 | **65.5** | 67.9 | 50.7 |
| | All* | 2098 | 44.7 | 75.0 | 49.2 | 66.1 | 55.9 | 59.1 | **63.5** | 60.5 | 71.1 | 74.1 | 57.0 | 60.5 | 61.4 | **60.9** |
| | HTPO | 2099 | 54.9 | 77.5 | 19.8 | 52.1 | 52.1 | 58.6 | 64.8 | 55.9 | 61.3 | 78.1 | **73.5** | 38.3 | 57.2 | 50.7 |
| SVM | All | 58213 | 50.6 | 77.0 | **61.6** | 66.1 | 48.5 | 71.2 | 47.8 | 48.9 | 86.9 | 87.3 | 47.3 | 54.9 | 61.2 | 46.7 |
| | All* | 5290 | 44.5 | 76.1 | 55.7 | **68.4** | 50.7 | 64.4 | **57.0** | **56.2** | 81.0 | 81.9 | **52.7** | 57.4 | 67.2 | 59.3 |
| | MHS | 5291 | **57.9** | **80.0** | 4.8 | 48.3 | 48.4 | **67.2** | 46.4 | 46.4 | 47.8 | 75.0 | 50.1 | 22.7 | 47.7 | 51.8 |
| All hate baseline | | | 29.0 | 25.0 | 49.0 | 9.0 | 8.0 | 24.0 | 29.0 | 11.0 | 19.0 | 20.0 | 9.0 | 44.0 | 23.0 | 10.0 |

Table 3: Macro-averaged F1 scores across all hate speech test sets and our manually annotated set (Indep). For each model, the table includes: (1) the performance of the model trained on all the datasets (All); (2) the performance of the model when trained on a balanced sample of all datasets of the same size as the best single-dataset baseline (All*); and (3) the best overall performing model trained on a single dataset (BERTweet: *MHS*, TimeLMs: *AYR*, RoBERTa: *AYR*, BERT: *HTPO*, SVM: *MHS*). The best result for each dataset and model is bolded.

| model | Train | sexism | racism | disability | sexual orientation | religion | other | not-hate | AVG |
|---|---|---|---|---|---|---|---|---|---|
| TimeLMs | All Datasets | **72.2** | **72.9** | **74.2** | **76.9** | **52.6** | **58.8** | **90.6** | **71.6** |
| | HateX | 52.1 | 16.5 | 0 | 58.8 | 31.8 | 5.8 | 86.0 | 35.9 |
| BERTweet | All Datasets | **73.1** | **72.5** | **74.1** | **77.6** | **48.6** | **59.3** | **90.9** | **70.9** |
| | HateX | 47.8 | 6.8 | 0 | 43.9 | 0 | 0 | 85.5 | 26.3 |
| RoBERTa | All Datasets | **70.4** | **72.4** | **73.9** | **76.5** | **47.3** | **55.5** | **90.3** | **69.5** |
| | HateX | 50.5 | 16.3 | 0 | 67.9 | 29.1 | 7.7 | 85.5 | 36.3 |
| BERT | All | **68.9** | **66.3** | **75.5** | **69.3** | **40.3** | **54.9** | 93.3 | **66.9** |
| | HateX | 40.4 | 16.0 | 0 | 66.2 | 15.9 | 0 | 85.4 | 32.0 |
| SVM | All | **62.7** | **67.0** | **71.5** | **70.5** | 4.1 | **49.0** | 59.11 | **81.9** |
| | HateX | 20.1 | 6.0 | 0 | 54.9 | **6.8** | 0 | 84.5 | 24.6 |
| Baseline (most frequent) | | 0 | 0 | 0 | 0 | 0 | 0 | 84.0 | 12.0 |

Table 4: F1 score for each class in the multiclass setting when trained on all the datasets (All) and when trained only with HateX. Macro-average F1 (AVG) is also reported.

by an expert. In total 200 tweets were annotated as hateful, not-hateful, or as "NA" in cases where the annotator was not sure whether a tweet contained hate speech or not. The *Indep* test set consists of 151 non-hate and 20 hate tweets and due to its nature (specific content & expert annotation) can be leveraged to perform a targeted evaluation on models trained on similar and unrelated data. While we acknowledge the limitations of the *Indep* test set (i.e., relative small number of tweets and only one annotator present), our aim is to use these tweets, collected using relatively simple guidelines[5], to test the overall generalisation ability of our models and how it aligns to what people think of hate speech.

[5]Annotator guidelines are available in Appendix A.

### 4.3 Results

#### 4.3.1 Binary Setting

Table 3 displays the macro-F1 scores achieved by the models across all test sets when fine-tuned: (1) on all available datasets (*All*), (2) on the best overall performing model trained on a single dataset, and (3) on a balanced sample of the unified dataset of the same data size as (2). When looking at the average performance of (1) and (2), it is clear that when utilising the combined data, all models perform considerably better overall. This increased performance may not be achieved across all the datasets tested, but it does provide evidence that the relatively limited scope of the individual datasets hinder the potential capabilities of our

models. An even bigger constrast is observed when considering the performance difference on the *DEAP* subset, which deals with a less common type of hate speech (prejudice towards Asian people), where even the best performing single dataset model achieves barely 19.79% F1 compared to the worst combined classifier with 49.27% F1 (BERT All / BERT *HTPO*).

To further explore the importance of the size and diversity of the training data we train and evaluate our models in an additional settings. Considering the sample size of the best performing dataset for each model, an equally sized training set is extracted from all available data while enforcing a balanced distribution between hate and not-hate tweets (*All\**). Finally, we make sure to sample proportionally across the available datasets. The results (Table 3) reveal the significance that a diverse dataset has in the models' performance. All models tested perform on average better when trained on the newly created subsets (*All\**) when compared to the respective models trained only on the best performing individual dataset. Interestingly, this setting also achieves the best overall scores on the *Indep.* set, which reinforces the importance of balancing the data. Nonetheless, all the transformers models still achieve their best score when trained on all the combined datasets (*All*) which suggests that even for these models, the amount of available training data remains an important factor of their performance.

### 4.3.2 Multiclass Setting

Similarly to our binary setting, utilising the combined datasets in the multiclass setting enhances the models' performance. As can be observed from Table 4, all the models struggle to function at a satisfactory degree when trained on the *HateX* subset only. In particular, when looking at the "disability" class, none of the models manage to classify any of the entries correctly. This occurs even though "disability" entries exist in the *HateX* training subset, albeit in a limited number (21). This behaviour suggests that even when information about a class is available in the training data, language models may fail to distinguish and utilise it. Imbalanced datasets are a common challenge in machine learning applications. This issue is also present in hate speech, in this case exacerbated given the nature of the problem (including a potential big overlap of features between classes) and the lack of resources available.

## 5 Analysis

In this section, we dissect the results presented in the previous section by performing a cross-dataset comparison and a qualitative error analysis.

### 5.1 Cross-dataset Analysis

Figure 1 presents a cross-dataset comparison of the language models used for the evaluation. The heatmap presents the results of the models fine-tuned and tested for all dataset pair combinations. All models evaluated tend to perform better when they are trained and tested on specific subsets (left diagonal line on the heat-maps). Even when we evaluate models on similar subsets, they tend to display a deterioration in performance. For example both *CMS* and *AYR* datasets deal with sexism but the models trained only on *CMS* perform poorly when evaluated on *AYR* (e.g. BERTweet-CSM achieves 87% F1 on *CSM*, but only 52% on *AYR*). Finally, it is observable again that the models trained on the combined datasets (column "all") display the best overall performance and attain consistently high results in each individual test set. When analysing the difficulty of each individual dataset when used as a test set, DEAP is clearly the most challenging one overall. This may be due to the scope of the dataset, dealing with East Asian Prejudice during the COVID-19 pandemic, which is probably not well captured in the rest of the datasets. When used as training sets, none of the individual datasets is widely generalisable, with the results of the model fine-tuned on them being over 10 points lower than when fine-tuned on the unified dataset in all cases.

### 5.2 Qualitative Error Analysis

Aiming to better understand the models' results we perform a qualitative analysis focusing on entries miss-classified by our best performing model, *TimeLMs-All*.

**Multiclass.** When considering the multiclass setting, common errors are tweets that have been labelled as hateful, e.g. "U right, probably some old n\*gga named Clyde" is labelled as *racism* and "@user @user she not a historian a jihadi is the correct term" as *religion*, but the model classifies them as *not-hate*. However, depending on the context and without having access to additional information (author/target of the tweet) these entries may not actually be hateful.
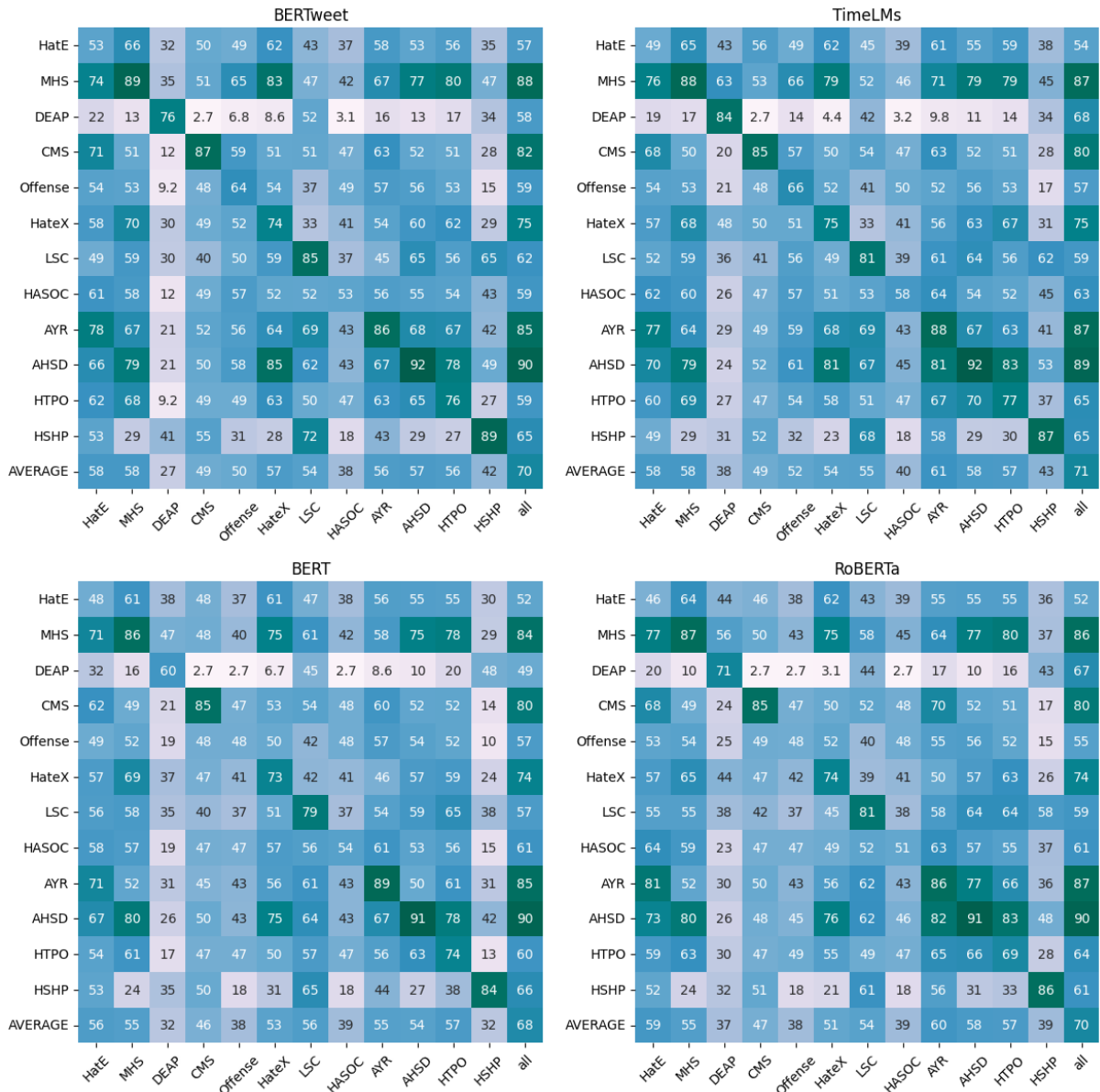
237

Figure 1: Macro-averaged F1 score for each dataset/model combination. The X axis indicates on which dataset the model was trained while the Y axis indicates the test set used to evaluate it. *AVERAGE* indicates the result by averaging across all datasets, and *all* represents the aggregated training set including all datasets.

It is also interesting to note the limitations that arise when training only on a single dataset, particularly if the data collection is done by utilising specific keywords. For example the tweets "Lana i love you b*tch. Put that flag back up h*e #lustfoflife" and "happy birthday b*tch, hope you have a great one h*e! @user" are correctly classified as *not-hate* by *TimeLMs-All* but are miss-classified as *sexism* by *TimeLMs-HateX*, despite *sexism* being present in the *HateX* dataset.

**Binary** In the binary setting, the model seems to struggle with entries such as "Meanwhile in Spain..#stopimmigration" and "This is outrageous.

Congress should be fired on the spot. #BuildThatWall #stopwastingmytaxdollars" where both entries are classified as *hate* but are labelled as *not-hate*. Similarly to the previous case, the classification of such tweets without additional context is a difficult task. While these tweets have hateful undertones, they may not be necessarily hate speech without considering them in their broader context.

Finally, when looking at the classification errors of *TimeLMs-AYR* (trained only on sexist and racist tweets) the need of diverse training data becomes apparent. For example, *TimeLM-AYR* fails to classify as hate speech the tweets "@user that r*tarded

guy should not be a reporter" and "I'm going to sell my iPhone and both my Macs, I don't support f*ggots." as hate speech in contrast to *TimeLMs-All* which classifies the tweets correctly as hateful.

# 6 Conclusion

In this paper, we presented a large-scale analysis of hate speech detection systems based on language models. In particular, our goal was to show the divergences across datasets and the importance of having access to a diverse and complete training set. Our results show how the combination of datasets make for a robust model performing competitively across all datasets. This is not a surprising finding given the size of the corresponding training sets, but the considerable gap (e.g. 70.7% to 61.0% in Macro-F1 for the best TimeLMs-21 performing model) shows that models trained on single datasets have considerable room for improvement. Moreover, even when controlling for data size, a model trained on a diverse set instead of a single dataset leads to better overall results.

As future work, we are planning to extend this analysis beyond English, in the line of previous multilingual approaches (Ousidhoum et al., 2019; Chiril et al., 2019; Bigoulaeva et al., 2021), and masked language models by including, among others, generative and instruction-tuning language models. In addition to the extensive binary-level evaluation, recognising the target group is a challenging area of research. While in Section 4.3.2, we provided some encouraging results, the results could be expanded with a unified taxonomy.

# 7 Ethics Statement

Our work aims to contribute and extend research regarding hate speech detection in social media and particular in Twitter. We believe that our efforts to contribute on the ongoing concerns around the status of hate speech on social medial.

We acknowledge the importance of the ACM Code of Ethics, and are committed on following it's guidelines. Our current work, uses either publicly available tweets under open licence and does not infringe any of the rules of Twitter's API. Moreover, given that our task includes user generated content we are committed to respect the privacy of the users, by replacing each user mention in the texts with a placeholder.

# 8 Limitations

In this paper, we have focused on existing datasets and a unification stemming from their features. The decisions taken to this unification, particularly in the selection of dataset and target groups, may influence the results of the paper.

We have focused on social media (particularly Twitter) and on the English language. While there has been extensive work on this medium and language, the conclusions that we can take from this study can be limiting, as the detection of hate speech involves other areas, domains and languages. In general, we studied a particular aspect of hate speech detection which may or not be generalizable.

Finally, due to computational limitations, all our experiments are based on base-sized language models. It is likely that larger models, while exhibiting similar behaviours, would lead to higher results overall.

# 9 Acknowledgements

# References

Raza Ali, Umar Farooq, Umair Arshad, Waseem Shahzad, and Mirza Omer Beg. 2022. Hate speech detection on twitter using transfer learning. *Computer Speech & Language*, 74:101365.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

James Bergstra, Daniel Yamins, and DD Cox. 2022. Hyperopt: Distributed asynchronous hyper-parameter optimization. *Astrophysics Source Code Library*, pages ascl–2205.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tulika Bose, Nikolaos Aletras, Irina Illina, and Dominique Fohr. 2022. Dynamically refined regularization for improving cross-corpora hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 372–382, Dublin, Ireland. Association for Computational Linguistics.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508, Dubrovnik, Croatia. Association for Computational Linguistics.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Gloria del Valle-Cano, Lara Quijano-Sánchez, Federico Liberatore, and Jesús Gómez. 2023. Socialhaterbert: A dichotomous approach for automatically detecting hate speech on twitter through textual analysis and user profiles. *Expert Systems with Applications*, 216:119446.

Fabio Del Vigna12, Andrea Cimino23, 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)*, pages 86–95.

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.

Ashkan Ebadi, Pengcheng Xi, Stéphane Tremblay, Bruce Spencer, Raman Pall, and Alexander Wong. 2021. Understanding the temporal evolution of covid-19 research through machine learning and natural language processing. *Scientometrics*, 126:725–739.

Mohamed Elsayed, Amira Abdelwahab, and Hatem Ahdelkader. 2019. A proposed framework for improving analysis of big unstructured data in social media. In *2019 14th International conference on computer engineering and systems (ICCES)*, pages 61–65. IEEE.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, and Showkat Hassan Malik. 2022. Detecting twitter hate speech in covid-19 era using machine learning and ensemble learning techniques. *International Journal of Information Management Data Insights*, 2(2):100120.

Lee Knuttila. 2011. User unknown: 4chan, anonymity and contingency. *First Monday*.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. 2018. Massively parallel hyperparameter tuning. *arXiv preprint arXiv:1810.05934*, 5.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 14–17.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224.

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

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

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.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

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.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Georg Rasch. 1960. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 573–584.

Andrew N Smith, Eileen Fischer, and Chen Yongjian. 2012. How does brand-related user-generated content differ across youtube, facebook, and twitter? *Journal of interactive marketing*, 26(2):102–113.

Collins Udanor and Chinatu C Anyanwu. 2019. Combating the challenges of social media hate speech in a polarized society: A twitter ego lexalytics approach. *Data Technologies and Applications*.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Samantha Walther and Andrew McCoy. 2021. Us extremism on telegram. *Perspectives on Terrorism*, 15(2):100–124.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

## A    Annotation Guidelines

In the following we present the guidelines provided to the annotator for the independent test set (Section 4.2).

A tweet is:

- labelled as "1" ("hate speech") if it contains any "discriminatory" (biased, bigoted or intolerant) or "pejorative" (prejudiced, contemptuous or demeaning) speech towards individuals or group of people.

- labelled as "0" ("not-hate-speech") if it does not contain hate speech as defined above.

- labelled "NA" if the coder is not sure whether the tweet contains hate speech or not.

The annotation should be based only on the text content of the tweet. This means that the coder should not follow any URL/media links if present.

## B    Hyperparameter Tuning

Table 5 lists the best hyperparameters for each of the models used in the evaluation.

| model | setting | learning rate | epochs | batch size | warm-up steps |
|---|---|---|---|---|---|
| TimeLMs | binary | 1.5857E-05 | 2 | 16 | 50 |
| BERTweet | binary | 1.4608E-05 | 2 | 4 | 100 |
| BERT | binary | 1.7882E-05 | 2 | 4 | 10 |
| RoBERTa | binary | 1.0377E-05 | 2 | 4 | 50 |
| TimeLMs | multiclass | 1.9100E-05 | 3 | 16 | 100 |
| BERTweet | multiclass | 9.0295E-06 | 3 | 4 | 10 |
| BERT | multiclass | 8.1260E-06 | 4 | 8 | 100 |
| RoBERTa | multiclass | 8.1260E-06 | 4 | 16 | 10 |

Table 5: Best hyper-parameters for models trained on the combined datasets for the binary and multiclass settings.