# MULTILINGUAL HATECHECK: Functional Tests for Multilingual Hate Speech Detection Models

**Paul Röttger**[1,2], **Haitham Seelawi**[1], **Debora Nozza**[1,3], **Zeerak Talat**[1,4], and **Bertie Vidgen**[1]

[1]Rewire    [2]University of Oxford    [3]Bocconi University    [4]Simon Fraser University

## Abstract

Hate speech detection models are typically evaluated on held-out test sets. However, this risks painting an incomplete and potentially misleading picture of model performance because of increasingly well-documented systematic gaps and biases in hate speech datasets. To enable more targeted diagnostic insights, recent research has thus introduced functional tests for hate speech detection models. However, these tests currently only exist for English-language content, which means that they cannot support the development of more effective models in other languages spoken by billions across the world. To help address this issue, we introduce MULTILINGUAL HATECHECK (MHC), a suite of functional tests for multilingual hate speech detection models. MHC covers 34 functionalities across ten languages, which is more languages than any other hate speech dataset. To illustrate MHC's utility, we train and test a high-performing multilingual hate speech detection model, and reveal critical model weaknesses for monolingual and cross-lingual applications.

## 1   Introduction

Hate speech detection models play a key role in online content moderation and also enable scientific analysis and monitoring of online hate. Traditionally, models have been evaluated by their performance on held-out test sets. However, this practice risks painting an incomplete and misleading picture of model quality. Hate speech datasets are prone to exhibit systematic gaps and biases due to how they are sampled (Wiegand et al., 2019; Vidgen and Derczynski, 2020; Poletto et al., 2021) and annotated (Talat, 2016; Davidson et al., 2019; Sap et al., 2021). Therefore, models may perform deceptively well by learning overly simplistic decision rules rather than encoding a generalisable understanding of the task (e.g. Niven and Kao, 2019; Geva et al., 2019; Shah et al., 2020). Further, aggregate and thus abstract performance metrics such as accuracy and F1 score may obscure more specific model weaknesses (Wu et al., 2019).

For these reasons, recent hate speech research has introduced novel test sets and methods that allow for a more targeted evaluation of model functionalities (Calabrese et al., 2021; Kirk et al., 2021; Mathew et al., 2021; Röttger et al., 2021b). However, these novel test sets, like most hate speech datasets so far, focus on English-language content. A lack of effective evaluation hinders the development of higher-quality hate speech detection models for other languages. As a consequence, billions of non-English speakers across the world are given less protection against online hate, and even the largest social media platforms have clear language gaps in their content moderation (Simonite, 2021; Marinescu, 2021).

As a step towards closing these language gaps, we introduce MULTILINGUAL HATECHECK (MHC), which extends the English HATECHECK functional test suite for hate speech detection models (Röttger et al., 2021b) to ten more languages. Functional testing evaluates models on sets of targeted test cases (Beizer, 1995). Ribeiro et al. (2020) first applied this idea to structured model evaluation in NLP, and Röttger et al. (2021b) used it to diagnose critical model weaknesses in English hate speech detection models. We create novel functional test suites for Arabic, Dutch, French, German, Hindi, Italian, Mandarin, Polish, Portuguese and Spanish.[1]  To our knowledge, MHC covers more languages than any other hate speech dataset.

The functional tests for each language in MHC broadly match those of the original HATECHECK, which were selected based on interviews with civil society stakeholders as well as a review of hate speech research. In each language, there are be-

---

[1]On dialects: we use Egyptian Arabic in Arabic script, European Dutch and French, High German, Standard Italian and Polish, Standard Hindi in Latin script, Standard Mandarin in Chinese script, Brazilian Portuguese and Argentinian Spanish.

tween 25 and 27 tests for different kinds of hate speech (e.g. dehumanisation and threatening language) as well as contrasting non-hate, which may lexically resemble hate speech but is clearly non-hateful (e.g. counter speech). These contrasts make the test suites particularly challenging to models that rely on overly simplistic decision rules and thus enable more accurate evaluation of model functionalities (Gardner et al., 2020). For each functional test, native-speaking language experts hand-crafted targeted test cases with clear gold standard labels, using the English cases as a starting point but adapting them to retain realism and cultural compatibility in the target language.

We demonstrate MHC's utility as a diagnostic tool by evaluating a multilingual XLM-T model (Barbieri et al., 2021) fine-tuned on a combination of three widely-used hate speech datasets in Spanish, Italian and Portuguese. This model achieves strong performance on the respective held-out test sets. However, testing with MHC reveals that the model is 1) overly sensitive to key words and key phrases, 2) biased in its target coverage and 3) error-prone and inconsistent in cross-lingual transfer, in both zero- and many-shot settings. If this model was used to moderate content, these critical weaknesses could cause serious harm, leaving some users unprotected from hate while others are restricted in their freedom of expression. We hope that by revealing such weaknesses, MHC can play a key role in the development of better multilingual hate speech detection models.[2]

**Definition of Hate Speech**  Following Röttger et al. (2021b), we define hate speech as *abuse that is targeted at a protected group or at its members for being a part of that group*. Protected groups are groups based on age, disability, gender identity, race, national or ethnic origins, religion, sex or sexual orientation, which broadly reflects Western legal consensus, particularly the US 1964 Civil Rights Act, the UK's 2010 Equality Act and the EU's Charter of Fundamental Rights. Based on these definitions, we approach hate speech detection as the binary classification of content as either hateful or non-hateful.

**Content Warning**  This article contains examples of hateful and abusive language. All examples are taken from MHC in their original language, to illustrate MHC's composition.

## 2   Constructing MHC

### 2.1   Defining Functional Tests

Functional tests in MHC test the ability of hate speech detection models to correctly classify different types of hateful and non-hateful content. The tests are fine-grained, distinct and contrastive. For example, we have separate tests for hate expressed using profanity (e.g. DE:"F*ck alle Juden"[3]) and non-hateful uses of profanity (e.g. DE:"Ich hatte noch nie so einen guten F*ck") in every language. Related functional tests are grouped into classes.

### 2.2   Selecting Functional Tests

We selected functional tests for each language in MHC to broadly match those from the original HATECHECK. Röttger et al. (2021b), in turn, motivated their selection of tests based on two factors: 1) a series of 21 interviews with NGO workers from the UK, US and Germany whose work directly relates to online hate, and 2) a review of previous hate speech research, particularly taxonomy work (e.g. Zampieri et al., 2019; Banko et al., 2020; Kurrek et al., 2020), error analyses (e.g. Davidson et al., 2017; van Aken et al., 2018; Vidgen et al., 2020) and survey articles (e.g. Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Vidgen et al., 2019). All test cases are short text statements, and they are constructed to be clearly hateful or non-hateful according to our definition of hate speech.

Overall, there are 27 functional tests grouped into 11 classes for each of the ten languages in MHC, except for Mandarin, which has 25 functional tests. Compared to the 29 functional tests in HATECHECK, we 1) exclude slur homonyms and reclaimed slurs, because they have no direct equivalents in most MHC languages, and 2) adapt functional tests for spelling variations to non-Latin script in Arabic and Mandarin. For Mandarin, there are two fewer tests for spelling variations and thus two fewer tests overall compared to the other nine languages. As in HATECHECK, the tests cover **distinct expressions of hate**, as well as **contrastive non-hate**, which shares lexical features with hate but is unambiguously non-hateful. We provide example cases in different languages for each functional test in Appendix A.

**Distinct Expressions of Hate**  MHC tests different types of derogatory hate speech (**F1-4**) and hate

---

expressed through threatening language (**F5/6**). It tests hate expressed using slurs (**F7**) and profanity (**F8**). MHC also tests hate expressed through pronoun reference (**F10/11**), negation (**F12)** and phrasing variants, specifically questions and opinions (**F14/15**). Lastly, MHC tests hate containing spelling variations such as missing characters or leet speak (**F23-34**), as well as spelling variations in non-Latin script for Arabic (**F28-31**) and Mandarin (**F32-34**). For example, there is an Arabic-specific test for spellings in Arabizi, the Arabic chat alphabet (**F30**), and a Mandarin-specific test for spellings in Pinyin, Mandarin's romanised version (**F34**).

**Contrastive Non-Hate**   MHC tests non-hateful contrasts which use profanity (**F9**) and negation (**F13**) as well as protected group identifiers (**F16/17**). It also tests non-hateful contrasts in which hate speech is quoted or referenced, specifically *counter speech*, i.e. direct responses to hate speech which seek to act against it (**F18/19**). Lastly, MHC tests non-hateful contrasts which target out-of-scope entities such as objects (**F20-22**) rather than a protected group.

## 2.3   Generating Test Cases

All test cases in MHC are hand-crafted by native-speaking language experts who have prior experience researching and/or annotating hate speech.[4] Each test case is a short statement that corresponds to exactly one gold standard label. HATECHECK's English test cases provide a starting point for MHC, but experts were encouraged to creatively adapt cases rather than providing literal translations, so as to retain relevance and realism. Adapting language-specific idioms (e.g. "murder that beat"), slurs (e.g. "c*nt") and profanity (e.g.  "f*ck") in particular required more creativity.

Test cases are generated at scale using templates (Dixon et al., 2018; Garg et al., 2019; Ribeiro et al., 2020), in which we replace tokens for protected group identifiers (e.g. NL:"Ik haat [IDENT].") and slurs (e.g. NL:"Voor mij ben je een [SLR]."). Compared to HATECHECK, the templates for MHC required more granular placeholders for gender- and case-inflected languages. German templates, for example, were instantiated based on gender and count of the identity group term, as well as its case: the male singular of Jew ("[male_IDENT_S]") in the German Akkusativ would be "den Juden" whereas its Nominativ would be "der Jude". The benefits

---

[4]See Appendix C for more detail on the language experts.

of the template approach are that 1) MHC has an equal number of cases targeted at different protected groups in each language, and 2) the templates can easily be used to generate more test cases targeted at other protected groups in the future.

MHC covers seven protected groups in each language. The selection of protected groups varies by language, with the goal of maximising the relevance of MHC to the cultural context of each language. Some groups, like women, are covered in all ten languages, whereas others, like lower caste people (HI:"shudra") are only covered in one language. For cases targeted at or referencing a particular protected group, we provide a secondary label for the group that is targeted. For an overview of protected groups covered by MHC in each language, see Appendix B.

In total, we generate 36,582 test cases across the ten languages in MHC, of which 25,511 (69.7%) are hateful and 11,071 (30.2%) are non-hateful. There is some variation in the number of test cases between different languages due to different numbers of functional tests and slurs, with Polish having the most cases (3,815) and Mandarin the least (3,378). 88.6% of all test cases are generated from templates. The remaining test cases do not use template tokens (e.g. PT:"O clima hoje está f*da") and are crafted individually. The average test case length is 8.5 words (std. dev. = 3.6) or 46.0 characters (std. dev. = 21.3). This resembles the composition of the original English HATECHECK.

**Validation**   To validate gold standard labels in MHC, all 36,582 test cases were reviewed by three native-speaking annotators. For each language we recruited at least 12 annotators, so that in total more than 120 annotators worked on MHC.[5] Annotation was prescriptive (Röttger et al., 2021a), meaning that annotators were tasked with adhering to clear annotation guidelines reflecting our definition of hate speech, rather than sharing their personal view on what is and is not hateful.[6] Compared to the original HATECHECK, where four out of five annotators confirmed the gold label for 99.4% of all test cases, there was more disagreement on MHC (see Appendix D). Upon review, language experts found that the vast majority of disagreement in MHC stemmed from annotators failing to apply the provided guidelines. For example, annotators mislabelled several instances of clearly hateful slurs

---

[5]See Appendix C for more detail on annotator recruitment.
[6]Annotation guidelines are available on GitHub.

(e.g. DE:"Du bist eine F*tze, mehr nicht"). There-fore, we did not exclude any cases from MHC. To enable further analysis and data filtering, we publish all annotator labels and mark up cases and templates where annotator majority labels differ from the gold labels.

## 3 Testing Models with MHC

### 3.1 Model Setup

As a suite of functional tests, MHC is broadly applicable across hate speech detection models for the ten languages that it covers. Users can test multilingual models across all ten languages or use a language-specific test suite to test monolingual models. MHC is model agnostic, and can be used to compare different architectures or different datasets in zero-, few- or many-shot settings, and even commercial models for which public information on architecture and training data is limited.

**Multilingual Transformer Models** We test XLM-T (Barbieri et al., 2021), an XLM-R model (Conneau et al., 2020) pre-trained on an additional 198 million Twitter posts in over 30 languages.[7] XLM-R is a widely-used architecture for multilingual language modelling, which has been shown to achieve near state-of-the-art performance on multilingual hate speech detection (Banerjee et al., 2021; Mandl et al., 2021). We chose XLM-T over XLM-R after initial experiments showed the former to outperform the latter on several hate speech detection datasets as well as MHC.

We fine-tune XLM-T on three widely-used hate speech datasets – one Spanish (Basile et al., 2019), one Italian (Sanguinetti et al., 2020) and one Portuguese (Fortuna et al., 2019). Accordingly, model performance is many-shot for Spanish, Italian and Portuguese, and zero-shot for all other languages.

All three datasets have an explicit label for hate speech that matches our definition of hate (§1), so that we can collapse all other labels into a single non-hateful label, to match MHC's binary format. The Spanish Basile et al. (2019) dataset contains 4,950 tweets, of which 41.5% are labelled as hateful. The Italian Sanguinetti et al. (2020) dataset contains 8,100 tweets, of which 41.8% are labelled as hateful. The Portuguese Fortuna et al. (2019) dataset contains 5,670 tweets, of which 31.5% are labelled as hateful.

We focus our discussion on XTC, an XLM-T model fine-tuned on a combination of these three datasets, which outperforms XLM-T models fine-tuned on the three datasets individually (see Appendix F). For the Spanish and Portuguese data, we use stratified 80/10/10 train/dev/test splits. For the Italian data, we use the original 91.6/8.4 train/test split, and then split the original training set into 90/10 train/dev portions. On the held-out test sets, XTC achieves 84.7 macro F1 for Spanish, 76.3 for Italian, and 73.3 for Portuguese, which is better than results reported in the original papers.[8]

**Testing Commercial Models** Few commercial models for hate speech detection are available for research use, and only a small subset of them can handle non-English language content. The best candidate for testing is Perspective, a free API built by Google's Jigsaw team.[9] Given an input text, Perspective provides percentage scores for attributes such as "toxicity" and "identity attack". The "toxicity" attribute covers a wide range of languages, including the ten in MHC. However, compared to hate speech, "toxicity" is a much broader concept, which includes other forms of abuse and profanity – some of which would be considered contrastive non-hate in the context of MHC. On the other hand, Perspective's "identity attack" aims to identify "negative or hateful comments targeting someone because of their identity" and thus aligns with our definition of hate speech (§1), but it is only available for three languages in MHC – German, Italian and Portuguese. For these three languages, XTC consistently outperforms Perspective (see Appendix H).

### 3.2 Results

**Performance Across Labels** MHC reveals clear gaps in XTC's performance across all ten languages (Table 2). Overall performance in terms of macro F1 is best on Mandarin (71.5), Italian (69.6) and Spanish (69.5), and worst on Hindi (58.1), Arabic (59.4) and Polish (66.2). F1 scores are higher for hateful cases than for non-hateful cases across all languages, with Hindi and Arabic exhibiting the biggest differences between hate and non-hate (∼40pp). For hateful cases, XTC performs best in terms of F1 score on Portuguese (83.5) and worst on Polish (76.1), but performance differences are

---

[7]We use the XLM-T implementation hosted on Hugging-Face: huggingface.co/cardiffnlp/twitter-xlm-roberta-base.

[8]See Appendix E for details on each dataset and pre-processing, and Appendix G for details on model training.
[9]www.perspectiveapi.com/

| | Functionality | Gold Label | Accuracy (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | AR | NL | FR | DE | HI | IT | ZH | PL | PT | ES |
| Derogation | **F1**: Expression of strong negative emotions (explicit) | hateful | 82.9 | 80.0 | 82.1 | 82.9 | 75.0 | 75.0 | 76.4 | 67.9 | **83.6** | 80.7 |
| | **F2**: Description using very negative attributes (explicit) | hateful | 87.9 | 87.1 | 84.3 | 82.1 | 77.1 | 82.1 | 85.7 | 72.9 | **91.4** | 78.6 |
| | **F3**: Dehumanisation (explicit) | hateful | 92.9 | 91.4 | 92.1 | 94.3 | 82.9 | 95.0 | **97.9** | 87.9 | 91.4 | 84.3 |
| | **F4**: Implicit derogation | hateful | 75.2 | 72.1 | 74.3 | 63.4 | **85.0** | 62.1 | 52.1 | 55.7 | 76.1 | 64.3 |
| Threat. language | **F5**: Direct threat | hateful | 85.0 | 94.3 | 92.9 | 93.6 | 84.3 | 88.6 | **97.9** | 76.4 | 90.0 | 89.3 |
| | **F6**: Threat as normative statement | hateful | 95.0 | 91.4 | 92.1 | 93.6 | **96.4** | 91.4 | 92.1 | 84.3 | 90.7 | 91.0 |
| Slurs | **F7**: Hate expressed using slur | hateful | 76.9 | 55.3 | 73.8 | 67.5 | 64.4 | 52.1 | 70.7 | 51.1 | **77.1** | *43.3* |
| Profanity usage | **F8**: Hate expressed using profanity | hateful | 94.3 | 83.6 | 91.0 | 90.0 | 80.0 | 80.0 | **97.1** | 79.3 | 94.3 | 75.7 |
| | **F9**: Non-hateful use of profanity | non-hate | 61.0 | 91.0 | 77.0 | 91.0 | 57.0 | 79.0 | 74.0 | 92.0 | 79.0 | **99.0** |
| Pronoun reference | **F10**: Hate expressed through reference in subsequent clauses | hateful | 84.3 | 81.4 | **94.0** | 89.3 | 90.7 | 91.4 | 83.6 | 65.0 | 86.4 | 84.1 |
| | **F11**: Hate expressed through reference in subsequent sentences | hateful | 88.6 | 90.0 | **91.4** | 91.4 | 89.3 | 87.9 | 89.3 | 69.3 | 85.0 | 79.3 |
| Negation | **F12**: Hate expressed using negated positive statement | hateful | **89.3** | 67.9 | 72.1 | 70.0 | 87.9 | 72.9 | 72.1 | 65.0 | 82.1 | 67.6 |
| | **F13**: Non-hate expressed using negated hateful statement | non-hate | *17.9* | *33.6* | *27.9* | *28.6* | *10.0* | *33.6* | *43.6* | *35.0* | *19.3* | *35.7* |
| Phrasing | **F14**: Hate phrased as a question | hateful | 88.6 | 73.6 | 84.3 | 91.4 | 77.9 | 87.9 | 74.3 | 75.0 | **93.6** | 72.9 |
| | **F15**: Hate phrased as an opinion | hateful | 90.7 | 77.9 | **92.9** | 87.9 | 78.6 | 89.3 | 90.0 | 75.0 | 92.1 | 78.6 |
| Non-hateful group identifier | **F16**: Neutral statements using protected group identifiers | non-hate | *44.3* | 67.1 | 67.1 | 67.9 | *49.3* | 83.6 | 80.0 | **88.6** | 56.6 | 68.6 |
| | **F17**: Positive statements using protected group identifiers | non-hate | *47.6* | 51.0 | 56.0 | 59.0 | *40.0* | **81.4** | 75.7 | 73.8 | 67.1 | 71.4 |
| Counter speech | **F18**: Denouncements of hate that quote it | non-hate | *10.3* | *47.6* | *22.8* | *31.6* | *17.1* | *37.9* | *43.9* | **61.4** | *22.4* | *40.9* |
| | **F19**: Denouncements of hate that make direct reference to it | non-hate | *9.6* | *32.9* | *18.6* | *34.8* | *13.7* | *31.7* | *45.1* | *44.9* | *24.8* | **64.6** |
| Abuse against non-protected targets | **F20**: Abuse targeted at objects | non-hate | 53.8 | 73.8 | 72.3 | 66.2 | *49.2* | 73.8 | 70.8 | 87.7 | 70.8 | **86.2** |
| | **F21**: Abuse targeted at individuals (not as member of a protected group) | non-hate | 56.2 | 66.2 | 72.3 | 67.7 | *41.5* | 60.0 | 52.3 | 90.8 | 55.4 | **92.3** |
| | **F22**: Abuse targeted at non-protected groups (e.g. professions) | non-hate | *27.7* | *47.7* | 55.4 | *33.8* | *26.2* | 56.9 | *38.5* | 61.5 | *44.6* | **70.8** |
| Spelling variations | **F23**: Swaps of adjacent characters | hateful | - | 82.1 | 89.3 | 89.3 | 87.9 | 75.7 | - | 69.3 | **94.3** | 84.3 |
| | **F24**: Missing characters | hateful | - | 85.0 | 75.0 | 82.9 | 74.3 | 69.3 | - | 72.9 | **85.7** | 68.6 |
| | **F25**: Missing word boundaries | hateful | - | 81.2 | 91.0 | 80.6 | 87.7 | 90.7 | - | 71.6 | **95.0** | 76.8 |
| | **F26**: Added spaces between chars | hateful | 65.8 | 61.2 | 86.8 | 85.8 | **89.7** | 77.0 | - | 58.0 | 83.2 | 71.3 |
| | **F27**: Leet speak spellings | hateful | - | 94.7 | **95.2** | 93.5 | 87.7 | 86.3 | - | 71.6 | 95.0 | 81.1 |
| | **F28**: AR: Latin char. replacement | hateful | **83.0** | - | - | - | - | - | - | - | - | - |
| | **F29**: AR: Repeated characters | hateful | 82.9 | - | - | - | - | - | - | - | - | - |
| | **F30**: AR: Arabizi (Arabic chat alphabet) | hateful | 60.9 | - | - | - | - | - | - | - | - | - |
| | **F31**: AR: Accepted alt. spellings | hateful | 85.6 | - | - | - | - | - | - | - | - | - |
| | **F32**: ZH: Homophone char. replacement | hateful | - | - | - | - | - | - | 89.3 | - | - | - |
| | **F33**: ZH: Character decomposition | hateful | - | - | - | - | - | - | 87.7 | - | - | - |
| | **F34**: ZH: Pinyin spelling | hateful | - | - | - | - | - | - | 76.5 | - | - | - |

Table 1: MHC covers 34 functionalities in 11 classes with a total of n = 36,582 test cases. 69.74% of cases (25,511 in 25 functional tests) are labelled hateful, 30.26% (11,071 in 9 functional tests) are labelled non-hateful. The right-most columns report accuracy (%) of the the XTC model (§3.1) across functional tests for each language. Languages which XTC was directly trained on are underlined, to highlight many-shot vs. zero-shot settings. Best performance on each functional test is **bolded**. Below random choice performance (<50%) is in *cursive red*. Examples of test cases for each functional test are listed in Appendix A.

relatively small across languages (< 8pp). For non-hateful cases, on the other hand, performance varies considerably across languages (< 24pp), with XTC performing best on Mandarin (61.1) and worst on Hindi (39.8).

| Language | F1-h | F1-nh | Mac. F1 |
|---|---|---|---|
| **Arabic / AR** | 79.1 | 39.8 | 59.4 |
| **Dutch / NL** | 80.1 | 53.3 | 66.7 |
| **French / FR** | 82.6 | 52.6 | 67.6 |
| **German / DE** | 82.6 | 55.2 | 68.9 |
| **Hindi / HI** | 78.5 | 37.7 | 58.1 |
| **Italian / IT** | 81.5 | 57.8 | 69.6 |
| **Mandarin / ZH** | 81.8 | 61.1 | 71.5 |
| **Polish / PL** | 76.1 | 56.4 | 66.2 |
| **Portuguese / PT** | 83.5 | 53.4 | 68.5 |
| **Spanish / ES** | 79.9 | 59.1 | 69.5 |

Table 2: Performance of XTC across the ten languages in MHC. Many-shot settings are underlined. All other languages are zero-shot. F1 score for **h**ateful and **n**on-**h**ateful cases, and overall macro F1 score.

**Performance Across Functional Tests**  Evaluating XTC on each functional test across languages (Table 1) reveals specific model weaknesses.

XTC performs better than a random binary-choice baseline (50% accuracy) on all functional tests for hate, with the exception of Spanish statements with hateful slurs (**F7**, 43.3% accuracy). Explicit dehumanisation (**F3**), threatening language (**F5/6**) and hate expressed using profanity (**F8**) appear to be the least challenging for the model, with relatively high and consistent accuracy across languages. In comparison, XTC generally performs worse on implicit hate (**F4**) and spelling variations (**F23+**). For other hateful functional tests, performance differs noticeably between languages. For example, XTC is very accurate on **F10**: hate expressed through reference in subsequent clauses in Spanish (94.0%), but much less so on Polish hate of the same kind (65.0%). Performance is worst on hate expressed using slurs (**F7**), with XTC misclassifying Spanish (43.3%), Polish (51.1%) and Italian statements (52.1%) at particularly high rates.

The only non-hateful functional test where performance is better than a random choice baseline across all ten languages is for non-hateful uses of profanity (**F9**). In contrast, negated hate (**F13**) and counter speech (**F18/19**), are misclassified at very high rates in every language. For example, XTC is

only 9.6% accurate on Arabic denouncements of hate that make direct reference to it (**F22**). Performance differs more noticeably across languages for non-hateful uses of protected group identity terms (**F16/17**) and abuse against non-protected targets (**F20-22**). For example, XTC is 81.4% accurate on positive statements using group identifiers in Italian, but only 40.0% accurate on the same type of statement in Hindi.

**Performance Across Target Groups**  MHC can show whether models exhibit *unintended biases* in target coverage (Dixon et al., 2018) by testing their performance on cases generated from [IDENT] templates. These cases are exactly the same, apart from the group that is targeted (e.g. IT: "Odio i gay" and IT:"Odio gli immigrati"). To illustrate, Table 3 shows model accuracy on these cases for French, Italian and Mandarin, which all cover the same seven protected groups.

| Target Group | FR | IT | ZH |
|---|---|---|---|
| Women | 70.9 | 70.1 | 74.5 |
| Trans people | 65.4 | 67.6 | 71.0 |
| Gay people | 56.4 | 68.2 | 70.3 |
| Black people | 61.5 | 67.7 | 66.7 |
| Disabled people | 61.2 | 64.4 | 72.6 |
| Muslims | 66.2 | 69.0 | 68.3 |
| Immigrants | 66.8 | 72.4 | 74.0 |

Table 3: Macro F1 of XTC on French, Italian and Mandarin test cases in MHC generated from [IDENT] templates, split by targeted protected group.

In French, XTC performs substantially worse on cases targeting gay people (56.4 macro F1) than on equivalent cases targeting women (70.9) or immigrants (66.8). In Italian, XTC performs particularly poorly on cases targeting disabled people (64.4), and in Mandarin on cases targeting black people (66.7), compared to other target groups.

### 3.3 Discussion

MHC reveals clear functional weaknesses in the XTC model across all ten languages.

First, XTC is overly sensitive to key words and phrases. XTC does have some ability to distinguish between different uses of potentially hateful key words, as demonstrated by its relatively good performance on both hateful and non-hateful uses of profanity across languages (**F8/9**). However, in languages like Hindi and Arabic, even entries

which refer positively to protected groups (**F17**) are misclassified as hateful at high rates. Accuracy on non-hateful contrasts to hateful phrases, like negated hate (**F13**) and counter speech (**F18/19**), is low across all languages. Thus, XTC appears to rely at least in part on overly simplistic decision rules (e.g. that the use of group identifiers is a signal for hate speech). It does not sufficiently register linguistic signals like negation, that reframe hateful phrases into clearly non-hateful statements (e.g. FR:"Je ne déteste pas les blacks.").

Second, XTC is biased in its target coverage across languages, classifying hate directed against some protected groups less accurately than equivalent cases directed at others (Table 3).

These are clear weaknesses that were not apparent from XTC's strong performance on held-out test sets. They match findings from HATECHECK for monolingual English models (Röttger et al., 2021b). If XTC was used to moderate content, these weaknesses could cause serious harm. In particular, misclassifying counter speech risks undermining positive efforts to fight hate speech, and biased target coverage may create and entrench biases in the protections afforded to different groups. However, the multilingual nature of MHC also allows for additional, novel insights.

First, we can evaluate cross-lingual performance in both zero- and many-shot settings (Table 2). XTC performs particularly well on Italian, Spanish and Portuguese – the languages it was fine-tuned on – but also on French, which is another Romance language. Performance on other European languages is also relatively high. By contrast, Hindi and Arabic clearly stand out as particularly challenging, with substantially lower performance. This suggests that cross-lingual transfer works better across more closely related languages and poses a challenge for more dissimilar languages.[10] Cultural differences across language settings may also affect transferability. We may for example expect hate in Italian and French to be more similar to each other than to hate in Hindi, along such dimensions as who the targets of hate are, which would likely affect the cross-lingual performance of hate speech detection models. Both hypotheses could be explored in future research.

Second, we can evaluate differences in language-specific model behaviour, again in zero- as well as many-shot settings. For example, XTC tends to overpredict hate in Hindi and Arabic, both zero-shot, whereas it tends to underpredict hate in many-shot Spanish and zero-shot Polish (Table 1). XTC also exhibits different target biases across languages, for zero-shot settings like in French and Mandarin as well as many-shot Italian (Table 3). This suggests that, in addition to accounting for differences in high-level performance, multilingual models may require very different calibration and adaptation across languages, even for languages they were not directly fine-tuned on.

Overall, the insights generated by MHC suggest two potential steps towards the development of more effective multilingual hate speech detection models: 1) creating training data in diverse languages to reduce language gaps, even for models with significant cross-lingual transfer abilities, and 2) evaluating and addressing language-specific model biases as well as differences in performance across languages.

## 4 Limitations

The limitations of the original HATECHECK also apply to MHC. First, MHC diagnoses specific model weaknesses rather than generalisable model strengths, and should be used to complement rather than substitute evaluation on held-out test sets of real-world hate speech. Second, MHC does not test functionalities related to context outside of individual documents or modalities other than text. Third, MHC only covers a limited set of protected groups and slurs across languages, but can easily be expanded using the provided case templates.

The multilingual nature of MHC creates additional considerations. First, comparisons of performance between languages are not strictly like-for-like, because cases in different languages are not literal translations of each other. This limitation is compounded for Arabic and Mandarin, which have unique functional tests for spelling variations. Second, even though MHC includes a diverse set of ten languages, these languages still only make up a fraction of languages spoken across the world. To our knowledge, MHC covers more languages than any other hate speech dataset, but hundreds of other languages remain neglected and should be considered for future expansions of MHC. Third, the selection of functional tests in MHC is based on HATECHECK, which was informed in part by

---

[10]The surprisingly good performance of XTC on Mandarin is a caveat, which may in part be explained by Mandarin being more prevalent than Arabic or Hindi in XLM-R's pre-training corpus (Conneau et al., 2020).

interviews in an anglo-centric setting. We worked with native-speaking language experts and created additional tests to account for non-Latin scripts in Arabic and Mandarin, but future research may consider additional interviews or other language-specific steps to inform expansions of MHC. Lastly, individual languages, like the ten included in MHC, are not monolithic but vary between speakers, especially across geographic regions and sociodemographic groups. We use widely-spoken dialects for the ten languages in MHC (see §1), but cannot cover all variations.

## 5   Related Work

**Diagnostic Hate Speech Datasets**   The concept of functional testing from software engineering (Beizer, 1995) was first applied to NLP model evaluation by Ribeiro et al. (2020). The original HATE-CHECK (Röttger et al., 2021b) then introduced functional tests for hate speech detection models, using hand-crafted test cases to diagnose model weaknesses on different kinds of hate and non-hate. Kirk et al. (2021) applied the same framework to emoji-based hate. Manerba and Tonelli (2021) provide smaller-scale functional test for abuse detection systems. Other research has instead collected real-world examples of hate and annotated them for more fine-grained labels, such as the hate target, to enable more comprehensive error analysis (e.g. Mathew et al., 2021; Vidgen et al., 2021). Instead of creating a static dataset, Calabrese et al. (2021) devise a hate speech-specific data augmentation technique based on simple heuristics to create additional test cases based on model training data. MHC is the first non-English diagnostic dataset for hate speech detection models.

**Non-English Hate Speech Data**   English is by far the most common language for hate speech datasets, as recent reviews by Vidgen and Derczynski (2020) and Poletto et al. (2021) confirm. Encouragingly, more and more non-English datasets are being created, particularly for shared tasks (e.g. Wiegand et al., 2018; Ptaszynski et al., 2019; Fersini et al., 2020; Zampieri et al., 2020; Mulki and Ghanem, 2021). However, very few datasets cover more than one language (Ousidhoum et al., 2019; Basile et al., 2019), and to our knowledge no dataset covers as many languages as MHC.

**Multilingual Hate Speech Detection**   The scarcity of non-English hate speech datasets has motivated research into few- and zero-shot cross-lingual hate speech detection, i.e. detection with little or no training data in the target language. However, model performance is generally found to be lacking in such settings (Stappen et al., 2020; Leite et al., 2020; Nozza, 2021). Others have thus explored data augmentation techniques based on machine translation, which yield limited improvements (Pamungkas et al., 2021; Wang and Banko, 2021). Overall, multilingual models trained or fine-tuned directly on the target languages, i.e. in many-shot settings, are still consistently found to perform best (Aluru et al., 2020; Pelicon et al., 2021). MHC's functional tests are model-agnostic and can be used to evaluate multilingual hate speech detection models trained on any amount of data.

## 6   Conclusion

In this article, we introduced MULTILINGUAL HATECHECK (MHC), a suite of functional tests for multilingual hate speech detection models. MHC expands the English-language HATECHECK (Röttger et al., 2021b) to ten additional languages: Arabic, Dutch, French, German, Hindi, Italian, Mandarin, Polish, Portuguese and Spanish. To our knowledge, MHC covers more languages than any other hate speech dataset. Across the languages, native-speaking language experts created 36,582 test cases, which provide contrasts between hateful and non-hateful content. This makes MHC challenging to hate speech detection models and allows for a more effective evaluation of model quality.

We demonstrated MHC's utility as a diagnostic tool by testing a high-performing multilingual transformer model, which was fine-tuned on three widely-used hate speech datasets in three different languages. MHC revealed the model to be 1) overly sensitive to key words and key phrases, 2) biased in its target coverage and 3) error-prone and inconsistent in cross-lingual transfer, in both zero- and many-shot settings.

So far, hate speech research has primarily focused on English-language content and thus neglected billions of non-English speakers across the world. We hope that MHC can contribute to closing this language gap and that by diagnosing specific model weaknesses across languages it can support the development of better multilingual hate speech detection models in the future.

## Acknowledgments

## References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, page 423–439, Berlin, Heidelberg. Springer-Verlag.

Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. *arXiv preprint arXiv:2111.13974*.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. XLM-T: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Boris Beizer. 1995. *Black-box testing: techniques for functional testing of software and systems*. John Wiley & Sons, Inc.

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.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Agostina Calabrese, Michele Bevilacqua, Björn Ross, Rocco Tripodi, and Roberto Navigli. 2021. Aaa: Fair evaluation for abuse detection systems wanted. In *13th ACM Web Science Conference 2021*, pages 243–252.

Arthur TE Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, et al. 2019. Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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 Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515. Association for the Advancement of Artificial Intelligence.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. Association for Computing Machinery.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA. Association for Computing Machinery.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, Tristan Thrush, and Scott A Hale. 2021. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *arXiv preprint arXiv:2108.05921*.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-grained fairness analysis of abusive language detection systems with CheckList. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91, Online. Association for Computational Linguistics.

Delia Marinescu. 2021. Facebook's content moderation language barrier. *New America*.

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.

Hala Mulki and Bilal Ghanem. 2021. Working notes of the workshop arabic misogyny identification (armi-2021). In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 7–8, New York, NY, USA. Association for Computing Machinery.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 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.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559. Publisher: PeerJ Inc.

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(2):477–523.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval 2019 Workshop*, page 89.

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.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021a. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021b. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Antonio Stranisci, Cristina Bosco, Caselli Tommaso, Viviana Patti, Irene Russo, et al. 2020. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–9. CEUR.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.

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.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Tom Simonite. 2021. Facebook is everywhere; its moderation is nowhere close. *Wired*.

Lukas Stappen, Fabian Brunn, and Björn W. Schuller. 2020. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. *CoRR*, abs/2004.13850.

Zeerak Talat. 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.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Talat, 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.

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.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Cindy Wang and Michele Banko. 2021. Practical Transformer-based Multilingual Text Classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 121–129, Online. Association for Computational Linguistics.

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, Melanie Siegel, and Josef Ruppenhofer. 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)*.

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.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, Florence, Italy. 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 SemEval*.

## A  Example MHC Test Cases

See Table 4 on the next page.

## B  Protected Groups in MHC

MHC covers seven protected groups per language.

**Arabic**: women /نساء, trans people / متحولين جنسيا, gay people / مثلين, black people / سود, Jews / يهود, disabled people / معاقين, immigrants / لاجئين

**Dutch**: women / vrouwen, trans people / transgenders, gay people / homoseksuelen, black people / zwarte personen, disabled people / gehandicapten, immigrants / immigranten, Muslims / Moslims

**French**: women / femmes, trans people / trans, gay people / homos, black people / blacks, disabled people / handicapés, Muslims / musulmans, refugees / réfugiés

**German**: women / Frauen, trans people / Transmenschen, gay people / Schwule, black people / Schwarze, disabled people / Behinderte Jews / Juden, refugees / Flüchtlinge

**Hindi**: women / aurato, trans people / hijro, gay people / gays, North-East Indians / assamese, Muslims / musalmano, Lower Caste / shudro, Black people / negro

**Italian**: women / donne, trans people / trans, gay people / gay, black people / persone di colore, disabled people / disabili, Muslims / musulmani, immigrants / immigrati

**Mandarin**: women / 女人, trans people / 跨性别者, gay people / 同性恋, black people / 黑人, disabled people / 残疾人, Muslims / 穆斯林, foreigners / 外国人

**Polish**: women / kobiety, trans people / osoby transpłciowe, gay people / geje, Asian people / azjaci, disabled people / niepełnosprawni, Jews / Żydzi, immigrants / imigranci

**Portuguese**: women / mulheres, black people / negros, gay people / gays, trans people / pessoas trans, indigenous people / indígenas, Jews / judeus, disabled people / deficientes

**Spanish**: women / mujeres, black people / negros, gay people / gays, trans people / trans, indigenous people / indígenas, Jews / judíos, disabled people / discapacitados

## C  Data Statement

Following Bender and Friedman (2018), we provide a data statement, which documents the generation and provenance of test cases in MHC.

**A. CURATION RATIONALE**   The goal of our research was to construct MHC, a multilingual suite of functional tests for hate speech detection models.  For this purpose, our team of native-speaking language experts generated a total of 36,582 short text documents in ten different languages, by hand and by using simple templates for group identifiers and slurs (§2.3). Each document corresponds to one functional test and a binary gold standard label (hateful or non-hateful).

**B. LANGUAGE VARIETY**   MHC covers ten languages: Arabic, Dutch, French, German, Hindi, Italian, Mandarin, Polish, Portuguese and Spanish.

**C. SPEAKER DEMOGRAPHICS**   All test cases across the ten languages in MHC were hand-crafted by native-speaking language experts – one per language.  All ten had previously worked on hate speech as researchers and/or annotators. Six out of ten experts identify as women, the rest as men. Four out of ten identify as non-White.

**D. ANNOTATOR DEMOGRAPHICS**   More than 120 annotators provided annotations on MHC, with at least 12 annotators per language. Annotators were recruited on Appen, a crowdworking provider. Appen gave no demographic information beyond guaranteeing that annotators were native speakers of the languages in which they completed their work. In setting up the annotation task and communicating with annotators, we followed guidance for protecting and monitoring annotator well-being provided by Vidgen et al. (2019).

**E. SPEECH SITUATION**   All test cases were created between November 2021 and January 2022.

**F. TEXT CHARACTERISTICS**   The composition of the dataset is described in detail in §2.2 and §2.3 of the article.

## D  Annotator Disagreement on MHC

Annotator disagreement on MHC (Table 5) is higher than on the original HATECHECK (Röttger et al., 2021b), where four out of five annotators agreed on the gold label in 99.4% of cases. There is a lot of variation in disagreement across languages, with most having less than 5% disagreement, and only Mandarin and French more than 10%. Upon review, our language experts found that the vast majority of disagreements stemmed from annotator error, where annotators failed to apply the explicit, prescriptive annotation guidelines they received.

| | Functionality | Gold Label | Language | Example |
|---|---|---|---|---|
| Derogation | **F1**: Expression of strong negative emotions (explicit) | hateful | FR | Il n'y a rien de plus horrible qu'une femme. |
| | **F2**: Description using very negative attributes (explicit) | hateful | HI | duniya mai sabse zyada murkh bas gays hi hote hain. |
| | **F3**: Dehumanisation (explicit) | hateful | NL | Gehandicapten zijn een plaag. |
| | **F4**: Implicit derogation | hateful | PL | Życie każdego owada jest cenniejsze niż życie kobiety. |
| Threat. language | **F5**: Direct threat | hateful | ZH | 只能给女人死刑才能解决问题。 |
| | **F6**: Threat as normative statement | hateful | ES | En un mundo justo, los negros como ustedes estarían todos muertos. |
| Slurs | **F7**: Hate expressed using slur | hateful | PT | Anormais simplesmente não conseguem ajudar a si mesmos. |
| Profanity usage | **F8**: Hate expressed using profanity | hateful | IT | Tipico di un trans essere un pezzo di merda. |
| | **F9**: Non-hateful use of profanity | non-hate | AR | لازم احاول قد مقدر ما بقاش حمار. |
| Pronoun reference | **F10**: Hate expressed through reference in subsequent clauses | hateful | HI | wo shudra insaan dikhte hai par hai wo parijeevi hi. |
| | **F11**: Hate expressed through reference in subsequent sentences | hateful | PT | É obvio que eles são estúpidios, afiinal são gays |
| Negation | **F12**: Hate expressed using negated positive statement | hateful | DE | Behinderte sind uns grundsätzlich nicht ebenbürtig. |
| | **F13**: Non-hate expressed using negated hateful statement | non-hate | PL | Nie stójmy bezczynnie, gdy niepełnosprawni są mordowani. |
| Phrasing | **F14**: Hate phrased as a question | hateful | IT | Sbaglio a pensare che questi disabili debbano essere condannati a morte? |
| | **F15**: Hate phrased as an opinion | hateful | ES | Si me preguntas, los negros como tú no son mejores que las ratas. |
| Non-hateful group identifier | **F16**: Neutral statements using protected group identifiers | non-hate | ZH | 我和两个跨性别者住在一起。 |
| | **F17**: Positive statements using protected group identifiers | non-hate | NL | Homoseksuelen horen thuis in de regering. |
| Counter speech | **F18**: Denouncements of hate that quote it | non-hate | AR | ايه اللي خلاك تقول انك بتتقرف من النساء. |
| | **F19**: Denouncements of hate that make direct reference to it | non-hate | HI | logo ko kinnar bulana bilkul theek nahi hai. |
| Abuse against non-protected targets | **F20**: Abuse targeted at objects | non-hate | IT | Questa serie di esercizi è così nauseante. |
| | **F21**: Abuse targeted at individuals (not as member of a protected group) | non-hate | PL | Nienawidzę cię. |
| | **F22**: Abuse targeted at non-protected groups (e.g. professions) | non-hate | DE | Ich kann Politiker wie dich wirklich nicht ausstehen. |
| Spelling variations | **F23**: Swaps of adjacent characters | hateful | NL | Zulke transgenders moeten worden geëxecueterd. |
| | **F24**: Missing characters | hateful | FR | Typique d'un homo d'être à ce point stupde. |
| | **F25**: Missing word boundaries | hateful | PT | Essas mulheresme dão tanto nojo. |
| | **F26**: Added spaces between chars | hateful | ES | Quiero m a t a r a todos los indígenas. |
| | **F27**: Leet speak spellings | hateful | HI | keval aurat hi 1tni badi murkh ho sakta hai. |
| | **F28**: AR: Latin char. replacement | hateful | AR | انا بكرهكم 5الص يا يهود. |
| | **F29**: AR: Repeated characters | hateful | AR | ما حد فيكو ينزل ضرررررب بالمثلين دول بالمطوة. |
| | **F30**: AR: Arabizi (Arabic chat alphabet) | hateful | AR | Alanwa3 dee mal el yahood lazm ytnafth feha el23dam. |
| | **F31**: AR: Accepted alt. spellings | hateful | AR | النساء لازمهم حبس في الآفاص. |
| | **F32**: ZH: Homophone char. replacement | hateful | ZH | 我想沙死所有黑人。 |
| | **F33**: ZH: Character decomposition | hateful | ZH | 这些外国人就该闭上他们的彳句嘴。 |
| | **F34**: ZH: Pinyin spelling | hateful | ZH | 所有女人都去si。 |

Table 4: Example test cases for each of the 34 functional tests in MHC. Examples were selected at random.

For example, hate and more general abuse were often confused, and abuse against non-protected targets was often labelled as hateful. Therefore, we did not exclude any cases from MHC. To enable further analysis and data filtering, we provide annotator labels with the test suite and mark up cases and templates where there is disagreement between the annotator majority labels and the gold labels from our language experts.

| Language | % Disagreement | n Disagreement |
|---|---|---|
| Arabic / AR | 7.05 | 252 |
| Dutch / NL | 9.61 | 362 |
| French / FR | 21.22 | 789 |
| German / DE | 4.20 | 153 |
| Hindi / HI | 4.88 | 174 |
| Italian / IT | 0.73 | 27 |
| Mandarin / ZH | 11.48 | 388 |
| Polish / PL | 2.04 | 78 |
| Portuguese / PT | 4.12 | 152 |
| Spanish / ES | 2.40 | 90 |

Table 5: Proportion of entries and absolute number of entries where at least 2/3 annotators disagreed with the expert gold label, for each language in MHC.

# E  Datasets for Model Fine-Tuning

## E.1  Sanguinetti et al. (2020) Italian Data

**Sampling**  The authors compiled 8,100 tweets sampled using keywords. 4,000 tweets come from HaSpeeDe 2018 (Bosco et al., 2018), which in turn originates in the Sanguinetti et al. (2018) dataset. The other 4,100 tweets were collected as part of the Italian hate speech monitoring project "Contro l'Odio" (Capozzi et al., 2019).

**Annotation**  The Sanguinetti et al. (2018) tweets were annotated in two phases, first by expert annotators, then by crowdworkers from CrowdFlower. Each tweet was annotated by two to three annotators for six attributes: *hate speech*, *aggressiveness*, *offensiveness*, *irony*, *stereotype*, and *intensity*. For inter-annotator agreement, the authors report a Krippendorff's Alpha of 38% for CrowdFlower, and a Cohen's Kappa of 45% for the expert annotators. The "Contro l'Odio" tweets were annotated by crowdworkers, but inter-annotator agreement was not reported. (Sanguinetti et al., 2020).

**Data**  We use all 8,100 tweets (41.8% hate).

**Definition of Hate Speech**  "Language that spreads, incites, promotes or justifies hatred or violence towards the given target, or a message that aims at dehumanizing, delegitimizing, hurting or

intimidating the target. The targets are Immigrants, Muslims, and Roma groups, or individual members of such groups."

## E.2  Fortuna et al. (2019) Portuguese Data

**Sampling**  Fortuna et al. (2019) initially collected 42,930 tweets based on a search of 29 user profiles, 19 keywords and ten hashtags. They then filtered the tweets, keeping only Portuguese-language tweets, and removing duplicates and retweets, resulting in 33,890 tweets. Finally, they set a cap of a maximum of 200 tweets per search method, to create the final dataset of 5,668 tweets.

**Annotation**  All tweets in the dataset were annotated as either hateful or non-hateful by 18 non-expert Portuguese native speakers were hired. Each tweet was annotated by three annotators, and inter-annotator agreement was low, with a Cohen's Kappa of 0.17.

**Data**  We use all 5,668 tweets (31.5% hate).

**Definition of Hate Speech**  "Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used."

## E.3  Basile et al. (2019) Spanish Data

**Sampling**  Tweets were sampled using three methods: 1) monitoring potential victims of hate accounts, 2) retrieving tweets from the history of identified haters, and 3) retrieving tweets using neutral and derogatory keywords, polarising hashtags, and stems. This yielded 19,600 tweets, of which 6,600 are in Spanish and the rest in English.

**Annotation**  The dataset was annotated for three attributes: *hate speech*, *target range* (individuals or groups), and *aggressiveness*. First, all data was annotated by at least three Figure Eight crowdworkers. Inter-annotator agreement on Spanish *hate speech* was high, with a Cohen's Kappa of 0.89. Second, two experts annotated each tweet. The final label was assigned based on majority vote across the crowd and expert annotators.

**Data**  We use all 6,600 Spanish tweets, of which 41.5% are labelled as hateful.

**Definition of Hate Speech** "Any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics."

### E.4 Pre-Processing

Before using the datasets for fine-tuning, we remove newline and tab characters. We replace URLs and user mentions with [URL] and [USER] tokens.

## F XLM-T Model Comparison

We denote the three XLM-T models trained on Italian Sanguinetti et al. (2020), Portuguese Fortuna et al. (2019) and Spanish Basile et al. (2019) as XLM-IT, XLM-PT and XLM-ES respectively. XTC denotes the XLM-T model trained on the combination of all three datasets, for which we report results in the main body of this article. XTC generally outperforms the monolingual models when compared on the respective held-out test sets (Table 6) as well as MHC (Table 7).

| Dataset | XLM-IT | XLM-PT | XLM-ES | XTC |
|---------|--------|--------|--------|------|
| IT | 73.2 | - | - | **76.3** |
| PT | - | **75.3** | - | 73.3 |
| ES | - | - | 84.0 | **84.7** |

Table 6: Macro F1 for each fine-tuned model on its respective test set and for XTC on all test sets.

| Lang. | XLM-IT | XLM-PT | XLM-ES | XTC |
|-------|--------|--------|--------|------|
| AR | 51.3 | 45.8 | 51.4 | **59.4** |
| NL | 59.9 | 49.5 | 59.6 | **66.7** |
| FR | 57.5 | 50.5 | 62.2 | **67.6** |
| DE | 62.1 | 46.9 | 59.5 | **68.9** |
| HI | 48.2 | 44.4 | 47.4 | **58.1** |
| IT | 53.6 | 47.0 | 54.6 | **69.6** |
| ZH | 61.8 | 42.7 | 53.2 | **71.5** |
| PL | 57.5 | 49.2 | 58.2 | **66.2** |
| PT | 58.6 | 64.2 | 56.0 | **68.5** |
| ES | 60.0 | 50.1 | 64.4 | **69.5** |

Table 7: Macro F1 across languages on MHC for each of our fine-tuned models.

## G XLM-T Model Details

**Model Architecture** We implemented XLM-T model (Barbieri et al., 2021) using the `transformers` Python library (Wolf et al., 2020). XLM-T is an XLM-R (Conneau et al., 2020) model pre-trained on an additional 198 million Twitter posts in over 30 languages. It has 12 layers,

a hidden layer size of 768, 12 attention heads and a total of 278 million parameters. For sequence classification, we added a linear layer with softmax output.

**Fine-Tuning** All models use unweighted cross-entropy loss and the AdamW optimiser (Loshchilov and Hutter, 2019) with a 5e-5 learning rate and a 0.01 weight decay. For regularisation, we set a 10% dropout probability, and for batch size we use 32. For each model, we train for 50 epochs, with an early stopping strategy with a patience of 5 epochs, with respect to improvements in the binary F1-score on the validation split. We store the checkpoint with the highest binary F1-score and use it as our final model.

**Computation** We ran all computations on an AWS "g4dn.2xlarge" server equipped with one NVIDIA T4 GPU card. The average wall time for each each training step was around 3 seconds.

**Model Access** We make the XTC model available for download on HuggingFace.

## H Google Perspective Results

We test Perspective's "identity attack" attribute and convert the percentage score to a binary label using a 50% cutoff. Testing was done in February 2022.

On the held-out test sets for Italian (Sanguinetti et al., 2020) and Portuguese (Fortuna et al., 2019), Perspective scored 70.7 and 64.1 macro F1. Perspective is outperformed on both languages by XTC, which scored 76.3 and 84.7 (Table 6).

On MHC, for the three languages it supports, Perspective (Table 8) performs worse than XTC (Table 2) in terms of macro F1 for Italian and Portuguese, and around equally well for German.

| Language | F1-h | F1-nh | Macro F1 |
|----------|------|-------|----------|
| **German / DE** | 84.1 | 54.9 | 69.5 |
| **Italian / IT** | 69.6 | 61.2 | 65.4 |
| **Portuguese / PT** | 84.2 | 47.6 | 65.9 |

Table 8: Performance of the Perspective API across the three languages it supports in MHC. F1 score for **h**ateful and **n**on-**h**ateful cases, and overall macro F1 score.