DoDo Learning: DOmain-DemOgraphic Transfer in Language Models for Detecting Abuse Targeted at Public Figures

Anonymous ACL submission

Abstract

Public figures receive disproportionate levels of abuse on social media, impacting their active participation in public life. Automated systems can identify abuse at scale but labelling training data is expensive and potentially harmful. So, it is desirable that systems are efficient and generalisable, handling shared and specific aspects of abuse. We explore the dynamics of cross-group text classification in order to understand how well models trained on one domain or demographic can transfer to others, with a view to building more generalisable abuse classifiers. We fine-tune language models to classify tweets targeted at public figures using our novel DoDo dataset, containing 28,000 entries with fine-grained labels, split equally across four DOmain-DemOgraphic pairs (male and female footballers and politicians). We find that (i) small amounts of diverse data are hugely beneficial to generalisation and adaptation; (ii) models transfer more easily across demographics but cross-domain models are more generalisable; (iii) some groups contribute more to generalisability than others; and (iv) dataset similarity is a signal of transferability.

Content Warning: We include some synthetic examples of the dataset schema to illustrate its contents.

Data Release Statement: Due to institutional guidelines concerning privacy issues surrounding the release of Twitter data, we are unable to release the DoDo dataset.

1 Introduction

Civil discussion between public figures and citizens is a key component of a well-functioning democratic society (Dewey, 1927; Rowe, 2015; Papacharissi, 2004). Social media has opened new channels of communication and permitted greater access between users and public figures (Doidge, 2015; Ward and McLoughlin, 2020); becoming an important tool for self-promotion, message spreading and maintaining a dialogue with fans, followers or the electorate (Farrington et al., 2014), beyond traditional media gatekeeping (Coleman, 1999, 2005; Coleman and Spiller, 2003; Williamson, 2009). However, there is a cost: the immediacy, ease and anonymity of online interactions has routinised the problem of abuse (Suler, 2004; Shulman, 2009; Brown, 2009; Joinson et al., 2009; Rowe, 2015; Ward and McLoughlin, 2020). Public figures attract more intrusive and abusive attention than average users of online platforms (Mullen et al., 2009; Meloy et al., 2008), and abuse directed towards them is both highly-public yet often grounded in highly-personal attacks (Erikson et al., 2021). There are detrimental effects to individual victims' mental health, which can ultimately result in their withdrawal from public life (Vidgen et al., 2021a; Delisle et al., 2019), and to society from normalising a culture of abuse and hate (Ingle, 2021). Disengagement is particularly worrisome for the functioning of democracy and political representation as it might be spread unevenly across groups (Theocharis et al., 2016; Greenwood et al., 2019; Ward and McLoughlin, 2020), e.g. women MPs being more likely to leave politics than men (Manning and Kemp, 2019).

Tackling abuse against public figures is a pressing issue, but the volume of social media posts makes manual investigations challenging, and conclusions drawn from anecdotal self-reporting or small sample size surveys offer limited and potentially biased coverage of the problem (Ward and McLoughlin, 2020). Automated systems based on machine learning or language models can be used to classify text at scale, but depend on labelling training data which is complex, expensive to collect and potentially psychologically harmful to annotators (Kirk et al., 2022c).

In this context, it is highly desirable to develop general abuse classifiers that can perform well across a range of different abuse types whilst being trained on a minimal 'labelling budget'. However, this may be technically challenging because, while some properties of abuse are shared across settings, different *domains* (e.g., sport, politics or journalism) introduce linguistic and distributional shifts. Furthermore, previous reports reveal that the nature of online abuse is heavily influenced by the identity attributes of its targets, for example gendered abuse against female politicians (Bardall, 2013; Stambolieva, 2017; Erikson et al., 2021; Delisle et al., 2019); so, learnings from different *demographics* may also not transfer. Exploring the effect of distributional shifts on model performance is useful for computational social scientists studying real-world phenomena, and for policymakers aiming to understand how to tackle online harm.

Despite the promise of generalisable abuse models for protecting more groups from harm, only a limited amount of research has addressed the extent to which models trained in one context can transfer to another. In this paper, we ask how well classifiers trained to detect abuse for one group transfer to others, with a view to building more generalisable models. Our novel DoDo dataset is collected from Twitter/X[1] and contains tweets targeted at public figures across two Domains (UK members of parliament or "MPs", and professional footballer players) and two Demographic groups (women and men). Tweets are annotated with four fine-grained labels to disambiguate abuse from other sentiments like criticism. We present results from experiments exploring the impacts of data diversity and number of training examples on domain-demographic transfer and generalisability.

2 Dataset

2.1 Data Collection

Our data is collected from Twitter. While generally over-researched (Vidgen and Derczynski, 2020), it is a dominant source for interactions between public figures and the general public. Most MPs have Twitter accounts and Twitter activity may even have a small impact on elections (Bright et al., 2020).

We compiled lists of accounts for UK MPs (590 accounts, 384 men, 206 women) and for players from England's top football divisions (808 from the Men's Premier League, 216 from the Women's Super League). We used the Twitter API Filtered Stream and Full Archive Search endpoints to collect all tweets that mention a public figure's account over a given time window.[2]

Levels of abusive content 'in-the-wild' are relatively low (Vidgen et al., 2019). In order to evaluate classifiers on realistic distributions while maximising their ability to detect abusive content, we randomly sample the test and validation datasets (preserving real-world class imbalance) but apply boosted sampling for the training dataset (ensuring the model sees enough instances of the rarer abusive class). We sample $7,000$ tweets in total for each domain-demographic pair: a $3,000$ train split, a $3,000$ test split, and a $1,000$ validation split.

Appendix C provides more detail on data collection, processing, and sampling.

2.2 Data Annotation

In the context of abuse detection, fine-grained labels can provide clarity for annotators, and enable more extensive error analysis, compared to binary labels. We employed annotators to label tweets with one of 4 classes of sentiment expressed towards public figures: Positive, Neutral, Critical, or Abusive, as defined below.[3]

1. **Positive:** Language that expresses support, praise, respect or encouragement towards an individual or group. It can praise specific skills, behaviours, or achievements, as well as encourage diversity and the representation of identities.
2. **Neutral:** Language with an unemotive tone or that does not fit the criteria of the other three categories, including factual statements, event descriptions, questions or objective remarks.
3. **Critical:** Language that makes a substantive negative assessment or claim about an individual or group. Negative assessment can be based on factors such as behaviour, performance, responsibilities, or actions, without being abusive.[4]
4. **Abusive:** Language containing threats, insults, derogatory remarks (e.g., hateful use of slurs and negative stereotypes), dehumanisation (e.g., comparing individuals to insects, animals, or trash), mockery, or belittlement towards an individual, group, or protected identity attribute (The Equality Act (2010)).

[1] Twitter has recently rebranded as "X". As the DoDo dataset was collected before the rebrand, we refer to the platform as Twitter exclusively.

[2] A similar approach is adopted in prior work that tracks public figure abuse (Gorrell et al., 2020; Ward and McLoughlin, 2020; Rheault et al., 2019).

[3] Labels are assigned based on the use of language, not the target of sentiment expressed.

[4] The annotator guidelines focused on distinguishing between abuse and criticism. Criticism must include a rationale for negative opinions on an individual's actions (not their identity)—it is not a form of "soft" abuse.

Split	Stance	dodo							
		fb-m		*fb-w*		*mp-m*		*mp-w*	
Train	Abusive	867	29%	481	16%	1007	34%	870	29%
	Critical	475	16%	282	9%	1283	43%	1353	45%
	Neutral	647	21%	719	24%	605	20%	628	21%
	Positive	1011	34%	1518	51%	105	3%	149	5%
Test	Abusive	103	3%	43	1%	392	13%	373	12%
	Critical	377	13%	89	3%	1467	49%	1471	49%
	Neutral	811	27%	767	26%	985	33%	927	31%
	Positive	1709	57%	2101	70%	156	5%	229	8%
Validation	Abusive	33	3%	14	1%	140	14%	135	13%
	Critical	93	9%	45	5%	484	48%	459	46%
	Neutral	335	34%	267	27%	332	33%	337	34%
	Positive	539	54%	674	67%	44	4%	69	7%
Random	Abusive	181	3%	75	1%	744	13%	661	12%
	Critical	642	12%	197	4%	2676	49%	2676	49%
	Neutral	1677	30%	1466	27%	1788	33%	1741	32%
	Positive	3000	55%	3762	68%	292	5%	422	7%

Table 1: Tweet counts across splits, dodos, and stances, with percentages within the dodo split. Includes counts and percentages for tweets from all splits selected by random sampling before annotation (5,500 tweets total per dodo).

The two domains were annotated sequentially in batches, but we updated our approach after the first batch as we found that crowdworkers struggled with the complexity of our task (see Appendix A for details). The final Cohen Kappa[5] for each domain was 0.50 for footballers and 0.67 for MPs.

2.3 Analysis

Terminology We abbreviate pairs of domain-demographic data as: fb-m (footballers-men), fb-w (footballers-women), mp-m (MPs-men), mp-w (MPs-women). We refer to any given domain-demographic pair as a dodo. We refer to groups of models that we train by the number of dodos included in the training data: `dodo1` for models trained using one domain-demographic pair, `dodo2` for models trained using two pairs, etc.

Overview The total dataset has 28,000 annotated entries, 7,000 for each dodo pair, with 3K/3K/1K test/train/validation splits. Table 1 shows class distributions across splits and counts of tweets sampled randomly pre-annotation.

Class Distributions The last row of Table 1 contains the randomly sampled entries across each dataset (ignoring keyword sampled entries which would skew the distributions). The majority of tweets in the MPs datasets are abusive or critical, in contrast to the footballers datasets where the majority class is positive, especially for fb-w. We also see slightly higher proportions of abusive tweets targeted at male demographic groups (fb-m, mp-m). Further analysis here is outside the scope of this paper, but it is notable how levels of abuse vary.

Tweet Length The MPs data contains longer tweets on average than the footballers data (125 vs. 84 characters), and has over twice as many tweets $\geq$ 250 characters (1,632 vs. 556 tweets). 62% of these longer ($\geq$250 characters) tweets for MPs are critical, implying the presence of detailed political debate.

3 Experiments

We conduct experiments to study how well model performance transfers across domains and demographics, and how the quantity and diversity of training data affects model generalisability. To reflect the focus on generalisability, we evaluate models on: (i) "seen" dodos (dodos used in training); (ii) "unseen" dodos (dodos not used in training)[6]; and (iii) the total evaluation set (including evaluation splits from all dodos). We train models on data from combinations of dodo pairs, and experiment with continued fine-tuning on the resulting models. We repeat experiments across 3 random seeds and 2 labelling budgets. We make predictions using the total test set (12,000), and calculate mean and standard deviation of Macro-F1 across the seeds. The Macro-F1 score represents a macro-average of per class F1 scores, neutralising class imbalance. We also investigate the correlation of Macro-F1 with dataset similarity.

Models We fine-tune deBERTa-v3 (**deBERT**, He et al., 2021)[7], using Huggingface's Transformers Library(Wolf et al., 2020). We used Tesla K80 GPUs through Microsoft Azure, training for 5 epochs with an early stopping patience of 2 epochs using Macro-F1 on the validation set, requiring a total of 155 GPU hours.

Dodo Combinations Our dataset has four dodo pairs, each with 3,000 training entries. There are 15 combinations of these pairs (if order does not matter): four single pairs (`dodo1`), six ways to pick two pairs (`dodo2`), four ways to pick three pairs (`dodo3`) and all pairs (`dodo4`). For all combinations, we randomly shuffle the concatenated training data before any training commences.

[5] Calculated using the generalised formula from Gwet (2014) to account for variable # of annotations per entry.

[6] All test sets are fully held out from training—by "seen" and "unseen" we only mean the domain or demographic.

[7] We also ran experiments on distilBERT (Sanh et al., 2019), but deBERTa-v3 had consistently higher performance, therefore we only present results for deBERTa-v3.

Labelling Budget For each training combination, we make two budget assumptions. In the **full budget** condition, we concatenate the training sets: 3,000 training entries for dodo1 experiments; 6,000 for dodo2 experiments; 9,000 for dodo3; and 12,000 for dodo4. In the **fixed budget** condition, we assume train budget is fixed at 3,000 entries and allocate ratios according to the dodo combinations: each included dodo makes up 100% of the budget for dodo1 experiments; 50% for dodo2; 33% for dodo3; and 25% for dodo4. This allows us to test the effects of training data composition without confounding effects of its size.

4 Results

4.1 Small amounts of diverse data are hugely beneficial to generalisable performance.

Table 2 provides an overview of the performance of models trained on all combinations of dodos. The increase in performance from adding data from new domains or demographics is not linear: the full budget dodo2 models only attain a one percentage point (pp) average increase in Macro-F1 Score for an additional 3,000 training entries. We also see the two dodo4 models are only separated by 3pp despite the full budget version being exposed to 4 times the amount of training data as the fixed budget version. This shows that gains from data diversity outweigh those from significantly greater quantities of data in training generalisable models.

4.2 Cross-demographic transfer is more effective than cross-domain.

Table 3 shows the comparisons for domain transfer and demographic transfer by Macro-F1 score on the seen and unseen portions of the test set, using the full-budget dodo2 models. For domain transfer, training on footballers gives a 0.654 F1 on the footballers dataset and 0.576 F1 on the MPs datasets. This is symmetric with training on MPs and testing on footballers. For demographic transfer, training on the male pairs and testing on female pairs faces no drop in performance. In contrast, training on women and testing on men leads to a small reduction in performance on the male data. In general, this demonstrates that transferring across domains is more challenging than transferring across demographics while keeping the domain fixed.

Model Group	**Train on** *fb-m*	*fb-w*	*mp-m*	*mp-w*	**Macro-F1** *Full*	*Fixed*
dodo1	✓				0.676	-
		✓			0.612	-
			✓		0.655	-
				✓	0.643	-
dodo2	✓	✓			0.667	0.673
			✓	✓	0.675	0.661
	✓		✓		0.723	**0.708**
		✓		✓	0.718	0.698
	✓			✓	0.722	**0.708**
		✓	✓		0.718	0.654
dodo3	✓	✓	✓		0.702	0.695
	✓	✓		✓	0.724	0.706
	✓		✓	✓	0.727	**0.708**
		✓	✓	✓	0.725	0.700
dodo4	✓	✓	✓	✓	**0.731**	0.701

Table 2: Table of Macro-F1 scores on the total test set for all possible training data combinations, in both full and fixed budget scenarios. Colour-coded according to increasing Macro-F1 Score, with best scores for each budget in bold.

Train on	**Test on** *Seen*		*Unseen*	
fb-m; fb-w	FBs	0.654	MPs	0.576
mp-m; mp-w	MPs	0.682	FBs	0.560
fb-m; mp-m	Men	0.718	Women	0.724
fb-w; mp-w	Women	0.722	Men	0.690

Table 3: Cross-domain and cross-demographic transfer with mean Macro-F1 for full-budget dodo2 models. We train on two dodos and evaluate on concatenated portions of the test set, e.g., we train *fb-w; fb-m* then test on *fb-w; fb-m* (seen) and *mp-m, mp-w* (unseen). Colour-coded according to increasing Macro-F1 Score.

4.3 Cross-domain models are more generalisable than cross-demographic.

Figure 1 shows that, as expected, performance on test sets from seen dodos is generally higher than on those from unseen dodos (we investigate exceptions in Appendix D.1). Within the dodo2 models, cross-demographic within-domain models (e.g., `fb-m;fb-w`) perform 10pp better on average on seen dodo evaluation sets than unseen ones, compared to a much narrower gap of 1pp on average for cross-domain models (e.g., `fb-w;mp-w`). We also see from Table 2 that cross-domain within-demographic dodo2 models outperform all cross-demographic within-domain dodo2 models on the total test set. This provides evidence that models trained on a single domain struggle to deal with out-of-domain examples, and that cross-domain models are more generalisable.

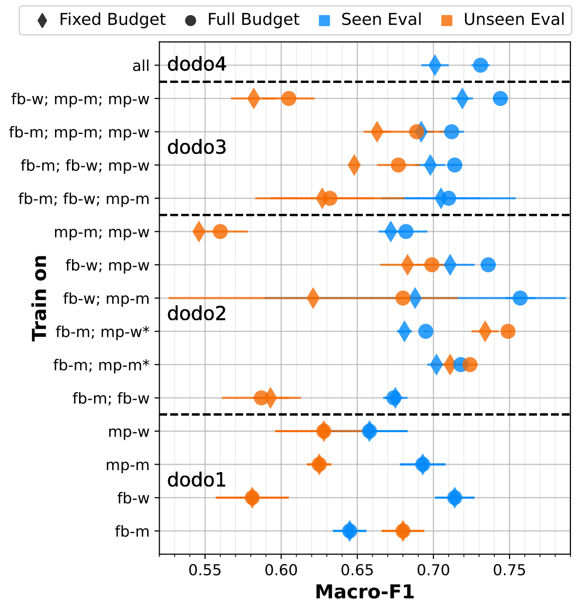

Figure 1: Mean and std-dev Macro-F1 across seeds for models trained on dodo combos, for fixed and full budgets, on test sets from seen and unseen dodos. *We removed one degenerate training seed (s=2).

4.4 Not all dodos contribute equally to generalisable performance.

The average Macro-F1 increase provided by including each dodo in training is summarised in Figure 2. `fb-m` provides the largest average increase in a fixed budget scenario, and `mp-w` in a full budget scenario.[8] In some cases, including fb-w data during training can detract from performance across both budgets. A dodo1 model trained only on fb-m also outperforms all other dodo1 models on the total test set (see Table 2), and fb-m data is included in the training dataset for the top ranking model for each dodo size across both labelling budgets. This suggests that training with fb-m is more important for good model generalisation than other dodos.

We now consider the situation of leaving out one dodo pair during training. We compare this left out case (dodo3) to training on all pairs (dodo4) in Table 4. We show the change in Macro-F1 on the total test set and change in number of training entries. For the full budget, leaving out mp-w from training leads to the largest reduction in performance. In contrast, removing all fb-w or mp-m entries does not significantly degrade performance even with 3,000 fewer training entries. For the fixed budget setting (with no confounding by training size), leaving out the two male pairs leads to a larger drop in performance than leaving out two female pairs.

[8]According to mean change in performance across all 7 possible scenarios of adding a dodo to training data.

	Raw size		Fixed size	
	Δ F1	**Δ N**	**Δ F1**	**Δ N**
all dodos	0.731	12,000	0.701	3,000
leave out fb-m	-0.006	-3,000	-0.001	0
leave out fb-w	-0.004	-3,000	0.007	0
leave out mp-m	-0.007	-3,000	0.005	0
leave out mp-w	-0.029	-3,000	-0.006	0

Table 4: Comparing model trained on all pairs (dodo4) with models trained on 3 pairs (dodo3). Shows relative change in mean Macro-F1 on total test set, and relative change in N of training entries.

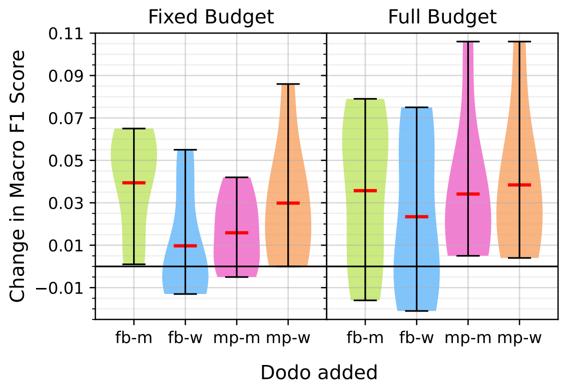

Figure 2: Violin plot displaying distribution of change in Macro-F1 score when adding a dodo to the training data (7 possible scenarios), with mean represented by red marker.

4.5 Only small amounts of data are needed to effectively adapt existing models to new domains and demographics.

Here we *start* with a fine-tuned specialist dodo1 model (i.e., a model fine-tuned on a single dodo) and *adapt* this model to a new dodo. We do continued fine-tuning of each fine-tuned dodo1 model on increments added from the adapt dodo train split.[9] For the models trained using each budget increment, we calculate Macro-F1 on test sets of both the start and adaption dodos (see Figure 3) so that we record both performance gains in adapting to new dodos alongside performance losses (forgetting) in seen dodos.

For almost all cases, the performance gain is notable after adding just 125 entries from the new dodo and increases with more entries. There is not a prominent performance gain after 500 entries except when adapting from fb-m to mp-m. This suggests that a small amount of data is efficient and cost-effective for testing how well existing models generalise. The importance of data composition

[9]The increments are [50, 125, 250, 500, 1000, 1500, 2000, 2500, 3000]. We train a separate model for each increment.

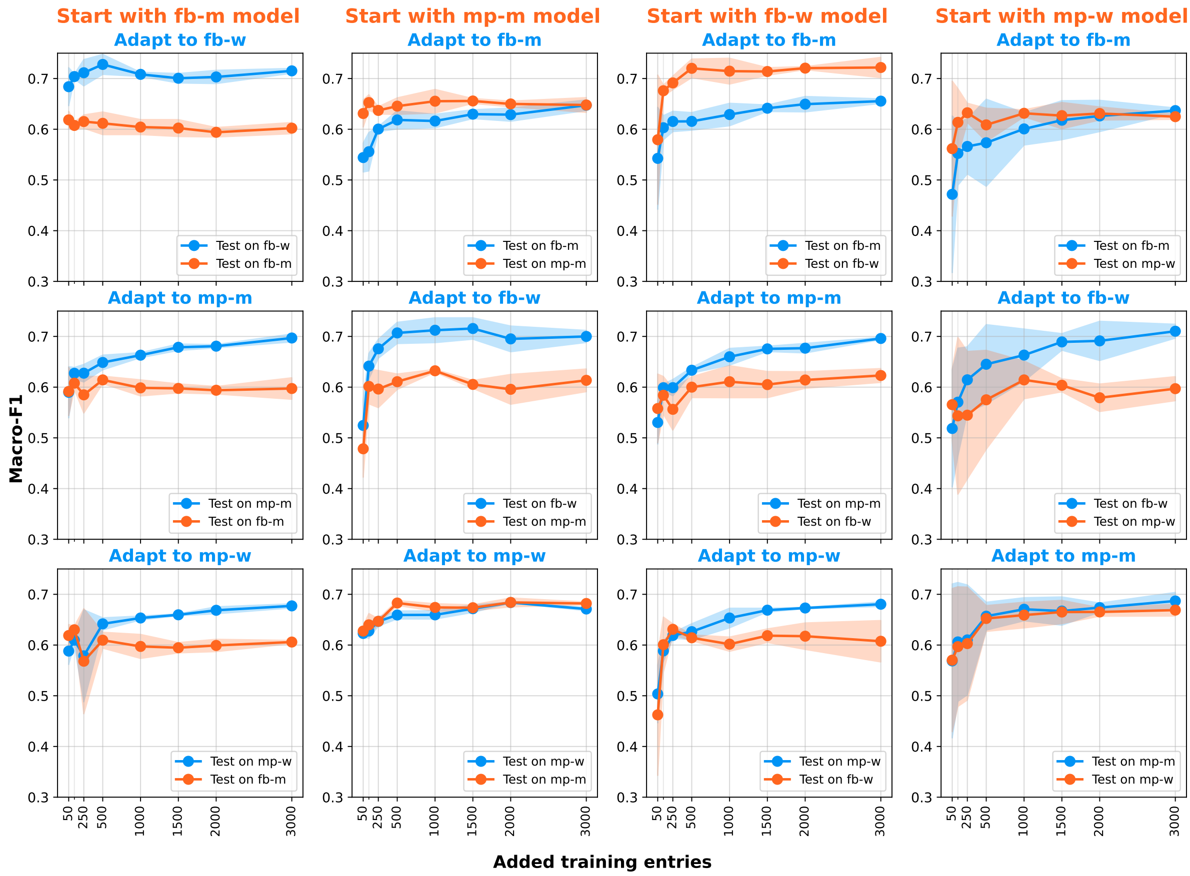

Figure 3: Learning curves for starting with a dodo1 model trained on a single dodo pair and adding increments from the training set of a new dodo pair. We show mean and std-dev Macro-F1 (across 3 seeds) on the new adapt dodo and source start dodo at each increment.

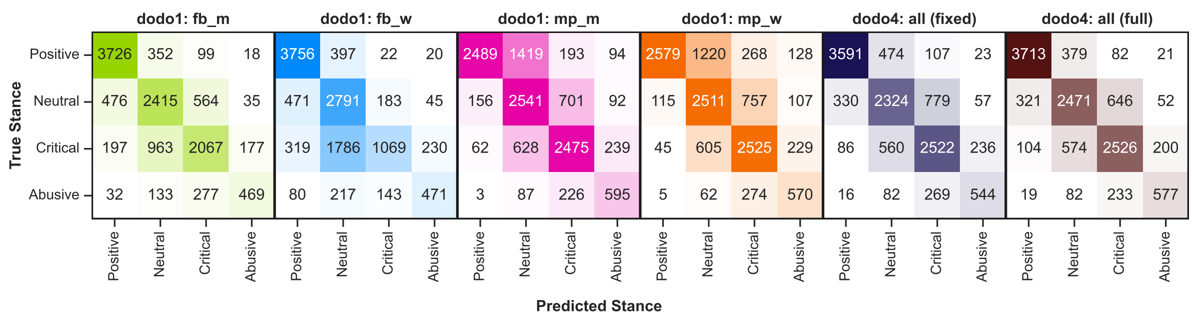

Figure 4: Confusion matrices for dodo1 and dodo4 models evaluated on the total test set (12,000 entries).

over data quantity aligns with the fixed/full budget findings from §4.1. On catastrophic forgetting, we generally do not find major performance drops. In some cases, adapting models to new data even helps classification in the source pair (e.g., mp-w to mp-m). Future work can explore where adaptation helps or hurts performance in source domains or demographics.

4.6 Dataset similarity is a signal of transferability.

Using the specialist dodo1 models, we examine if dataset similarity signals transferability, i.e., the Macro-F1 score that a dodo1 model can achieve on unseen dodos. We compute three classical text distance metrics with unigram bag-of-words approaches: Jaccard and Sørensen-Dice similarity, and Kullback-Leibler divergence. In Figure 5, we plot Macro-F1 scores (of unseen single dodos) against Jaccard similarity for each pair of dodos. The correlation coefficient is 0.7, demonstrating a positive relationship between dataset similarity and unseen dodo performance.[10] Greater similarity between demographic pairs versus domain pairs

[10]Correlation coefficients are 0.7 for Dice Similarity and -0.66 for KL Divergence, confirming Jaccard robustness.

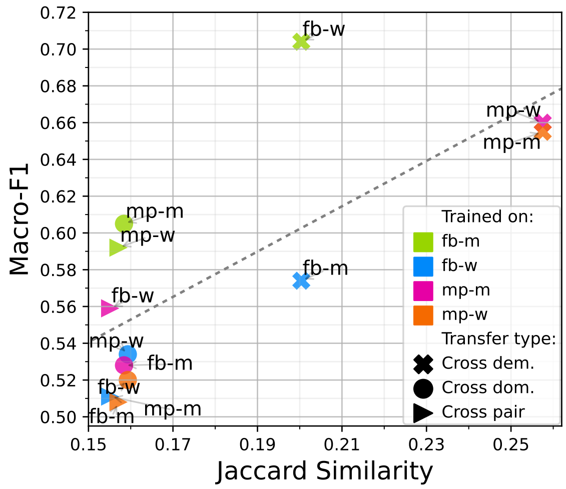

Figure 5: Jaccard similarity and mean 0-shot Macro-F1 for dodo1 deBERT models with line of best fit. On graph annotations represent evaluation dodo. Shows positive correlation ($\rho = 0.7$) and effectiveness of cross-demographic vs. cross-domain transfer.

results in better cross-demographic transfer versus cross-domain transfer. Using these metrics could help estimate transfer potential before investing in an expensive labelling process.

4.7 Error Analysis

We find that errors made by dodo1 models reflect the class imbalances outlined in Section 2.3. We also see errors relating to inherent similarities across bordering classes, demonstrating the value of fine-grained labels. We present confusion matrices on the total test in Figure 4, and full error analysis in Appendix D.2.

5 Discussion

We discuss the limitations of this work in Section 9, addressing difficulties in disentangling the direction of sentiment in social media posts, limitations in the chosen label schema, and the consequences of the chosen evaluation approaches. Here, we present avenues for future work.

Expanding demographics and adding more complexity to the labelling schema would provide a broader basis for understanding generalisability in abuse classification. Other promising avenues include investigating whether active learning techniques (Vidgen et al., 2022; Kirk et al., 2022c) aid more efficient cross-domain/demographic transfer, or whether architectures better suited for continual learning can assist in the addition of new groups without forgetting those previously trained-on (Hu et al., 2020; Qian et al., 2021; Li et al., 2022). We shuffled entries during training and used all four class labels but future work could assess whether performance is affected by order of training on different groups, and the impact of training on binary versus multi-class labels on transfer performance. Finally, our experiments only use fine-tuning on labelled data, but in-domain continued pre-training could be explored as a budget-efficient way to boost performance (Gururangan et al., 2020; Kirk et al., 2023).

6 Related Works

Abuse Against MPs Academics and journalists account abuse against politicians, which may cause politicians to withdraw from their posts (Parliament, 2023; Manning and Kemp, 2019; James et al., 2016). Empirical work commonly studies Twitter (Binns and Bateman, 2018; Gorrell et al., 2020; Ward and McLoughlin, 2020; Agarwal et al., 2021), including across national contexts such as European Parliament elections (Theocharis et al., 2016), Canadian and US politicians (Rheault et al., 2019) and members of the UK parliament (Gorrell et al., 2020). Other studies focus on gender differences in abuse (Rheault et al., 2019; Erikson et al., 2021) though some datasets only contain abuse against women (Stambolieva, 2017; Delisle et al., 2019) which limits comparison across genders (unlike DoDo). Various techniques are employed to identify abusive tweets including rules-based or lexicon approaches and topic analysis (Gorrell et al., 2018, 2020; Greenwood et al., 2019); traditional machine learning classifiers (Stambolieva, 2017; Rheault et al., 2019; Agarwal et al., 2021) or pre-trained language models and off-the-shelf classifiers like Perspective API (Delisle et al., 2019).

Abuse Against Footballers Sport presents a good case for studying public figure abuse due to the influence of athletes (Carrington, 2012), as well as the heightened symbolic focus on in-out groups and race-nation relations (Bromberger, 1995; Back et al., 2001; King, 2003; Burdsey, 2011; Doidge, 2015). Several studies track the change from racist chants at football stadiums, to the more pernicious and harder to control online abuse (King, 2004; Cleland, 2013; Cleland and Cashmore, 2014; Kilvington and Price, 2019). Civil society organisations track social media abuse as far back as the 2012/2013 season, but are limited by a focus on manual case-by-case resolution and suffer from chronic underreporting (Bennett and Jönsson, 2017). We build on our previous work in [REDACTED], which presents some of the same data as the male footballers portion in DoDo but also labels additional data using active learning.

Abuse Datasets and Detection Developing robust abuse classifiers is challenging (Zhang and Luo, 2019). Surveys on abuse detection cover various aspects such as algorithms (Schmidt and Wiegand, 2017; Mishra et al., 2019), model generalisability (Yin and Zubiaga, 2021), and data desiderata (Vidgen and Derczynski, 2020). Many studies curate data from mainstream platforms, focusing on abuse against different identities such as women (Fersini et al., 2018; Pamungkas et al., 2020) and immigrants (Basile et al., 2019). Recent approaches to developing abuse classifiers predominately fine-tune large language models on labelled datasets directly (Fortuna et al., 2021) (our approach) or in a multi-task setting (Talat et al., 2018; Yuan and Rizoiu, 2022), as well as incorporate contextual information (Chiril et al., 2022). Abuse detection datasets mostly focus on binary classification, and few cast the predictions as a multi-class problem. Similarly, cross-domain classification is under-explored to address generalisability (Glavaš et al., 2020; Yadav et al., 2023). The dataset we use in this paper rectifies some of these issues, as a cross-domain and demographic dataset with fine-grained labels.

Domain Adaptation Several NLP techniques have been explored for model generalisation in abuse detection, including feature-based domain alignment (Bashar et al., 2021; Ludwig et al., 2022), regularisation methods (Ludwig et al., 2022), and adaptive pre-training (Faal et al., 2021). Systematic evaluation of model generalisability is limited, focusing on dataset features (Fortuna et al., 2021) and multilinguality (Pamungkas et al., 2020; Yadav et al., 2023). To our knowledge, no prior work has quantified the effects of cross-domain *and* cross-demographic transfer.

7 Conclusion

We fine-tuned language models using our new DoDo dataset to classify abuse targeted at public figures for two domains (sports, politics) and two demographics (women, men). We found that (i) even small amounts of diverse data provide significant benefits to generalisable performance and model adaptation; (ii) cross-demographic transfer (from women to men, or vice-versa) is more effective than cross-domain transfer (from footballers to MPs, or vice-versa) but models trained on data from one domain are less generalisable than models trained on cross-domain data; (iii) not all domains and demographics contribute equally to training generalisable models; and (iv) dataset similarity is a signal of transferability.

There are broader policy implications of our work. Policymakers, NGOs and others with an interest in independently monitoring harms face challenges in building models that are both broad enough to capture a wide range of harms but also specific enough to capture the distinctive nature of abuse (e.g., the difference between hate speech targeted at male and female MPs); while remaining within resource constraints typical of policy settings. Our work contributes to the policy landscape by bringing fresh perspective on the feasibility of transferring models created to detect harm for one target to other targets. It thus provides insight into developing automated systems that are cost-effective, generalisable and performative across domains and demographics of interest.

8 Ethics and Harm Statement

We present our limitations section in §9. In addition to these limitations, engaging with a subject such as online abuse raises ethical concerns. Here we set out the nature of those concerns, and how we managed them. Creation and annotation of a dataset focusing on abuse risks harming the annotators and researchers constructing the dataset, as repeated exposure to such material can be detrimental towards their mental health (Kirk et al., 2022a). Mitigating these risks is easier with a small trained team of annotators (like those we used for the MPs datasets) and harder with crowdworkers (like those we used for the footballers datasets). With the trained group of annotators, we maintained an open annotator forum where they could discuss such issues during the labelling process, and seek welfare support. For crowdworkers, we had very limited contact with them but include on our guidelines and task description extensive content warnings and links to publicly-available resources on vicarious trauma.

We acknowledge that all experiments and data collection protocols are approved by the internal ethics review board at our institution.

9 Limitations

Targets of Abuse It is sometimes hard to disentangle the target of sentiment in tweets directed at public figures—some tweets praise public figures while simultaneously criticising another figure or even abusing identity groups (such as an praising an MP's anti-immigration policy while abusing immigrants). Our label schema does not tag target-specific spans nor flag when it is a non-public figure account or abstract group is being abused. We also do not use further conversational context during annotation. Furthermore, we are limited by gender distinctions in UK MPs statistics and football leagues—the dataset does not cover non-binary identities or other identity attributes.

Types of Abuse While our dataset is more diverse than most abuse datasets in including four class labels, it does not disaggregate abusive content into further subcategories such as identity attacks. Our preliminary keyword analysis suggested that identity attacks comprise a relatively small proportion of all abuse (especially for female footballers) but can nonetheless cause significant harm (Gelber and McNamara, 2016). Further investigation on abuse across demographic groups is needed to understand how women and men are targeted differently, and to assess distributional shifts of specific homophobic, racist, sexist or otherwise identity-based abuse.

Language and Platform Focus Our dataset contains English language tweets associated with UK MPs and the top football leagues in England (though players come from a variety of nationalities). Prior studies suggest politicians face online abuse in other countries (Theocharis et al., 2016; Ezeibe and Ikeanyibe, 2017; Rheault et al., 2019; Fuchs and Schäfer, 2020; Erikson et al., 2021); and that the English football social media audience is a global one (Kilvington and Price, 2019). However, shifting national or cultural context will introduce further distributional and linguistic shifts. Furthermore, our data is only collected from Twitter though abuse towards public figures exists on a variety of social media platforms (Agarwal et al., 2021) such as YouTube (Esposito and Zollo, 2021) or WhatsApp (Saha et al., 2021).

Evaluation Approach Aggregate evaluation metrics may obscure per dodo and per class weaknesses (Röttger et al., 2021). The Macro-F1 score across the combined test set from all dodos does not equal the averaged Macro-F1 across each dodo test set (the former is 4.7pp higher on average). This is due to different class distributions across dodos skewing the total Macro-F1 calculation. The ranking of models was consistent across these two metrics. We have not investigated the relative dataset difficulty (Ethayarajh et al., 2022) of individual dodo test sets, which may influence measures of generalisibility.

References

2010. Equality Act 2010. https://www.legislation.gov.uk/ukpga/2010/15/contents.

Pushkal Agarwal, Oliver Hawkins, Margarita Amaxopoulou, Noel Dempsey, Nishanth Sastry, and Edward Wood. 2021. Hate Speech in Political Discourse: A Case Study of UK MPs on Twitter. *Proceedings of the 32st ACM Conference on Hypertext and Social Media*, pages 5–16. Publisher: ACM.

Les Back, Tim Crabbe, John, and John Solomos Solomos. 2001. *The Changing Face of Football. Racism, Identity and Multiculturc in the English Game.* Berg Publishers.

Gabrielle Bardall. 2013. Gender-Specific Election Violence: The Role of Information and Communication Technologies. *Stability: International Journal of Security & Development*, 2(3):60.

Md Abul Bashar, Richi Nayak, Khanh Luong, and Thirunavukarasu Balasubramaniam. 2021. Progressive domain adaptation for detecting hate speech on social media with small training set and its application to covid-19 concerned posts. *Social Network Analysis and Mining*, 11:1–18.

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.

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.

Hayley Bennett and Anna Jönsson. 2017. Klick it out: tackling online discrimination in football. In *Sport and Discrimination*, page 12. Routledge.

Amy Binns and Martin Bateman. 2018. And they thought Papers were Rude. *British Journalism Review*, 29(4):39–44.

Jonathan Bright, Scott Hale, Bharath Ganesh, Andrew Bulovsky, Helen Margetts, and Phil Howard. 2020. Does campaigning on social media make a difference? Evidence from candidate use of Twitter during the 2015 and 2017 U.K. elections. *Communication Research*, 47(7):988–1009.

Christian Bromberger. 1995. Football as world-view and as ritual. *French Cultural Studies*, 6(18):293–311.

Christopher Brown. 2009. WWW.HATE.COM: White Supremacist Discourse on the Internet and the Construction of Whiteness Ideology. *Howard Journal of Communications*, 20(2):189–208.

Daniel Burdsey. 2011. *Race, Ethnicity and Football*. Routledge.

Ben Carrington. 2012. Introduction: sport matters. *Ethnic and Racial Studies*, 35(6):961–970.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, pages 1–31.

Jamie Cleland. 2013. Racism, Football Fans, and Online Message Boards. *Journal of Sport and Social Issues*, 38(5):415–431. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

Jamie Cleland and Ellis Cashmore. 2014. Fans, Racism and British Football in the Twenty-First Century: The Existence of a 'Colour-Blind' Ideology. *Journal of Ethnic and Migration Studies*, 40(4):638–654.

Stephen Coleman. 1999. Can the New Media Invigorate Democracy? *The Political Quarterly*, 70(1):16–22.

Stephen Coleman. 2005. New mediation and direct representation: reconceptualizing representation in the digital age. *New Media & Society*, 7(2):177–198.

Stephen Coleman and Josephine Spiller. 2003. Exploring new media effects on representative democracy. *The Journal of Legislative Studies*, 9(3):1–16.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pages 512–515. Publisher: AAAI Press.

Laure Delisle, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin, and Julien Cornebise. 2019. A large-scale crowdsourced analysis of abuse against women journalists and politicians on twitter.

John Dewey. 1927. *The public and its problem*. Henry Holt.

Mark Doidge. 2015. 'If you jump up and down, Balotelli dies': Racism and player abuse in Italian football. *International Review for the Sociology of Sport*, 50(3):249–264. Publisher: SAGE PublicationssSage UK: London, England.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 42–51.

Josefina Erikson, Sandra Håkansson, and Cecilia Josefsson. 2021. Three Dimensions of Gendered Online Abuse: Analyzing Swedish MPs' Experiences of Social Media. *Perspectives on Politics*, pages 1–17. Publisher: Cambridge University Press.

Eleonora Esposito and Sole Alba Zollo. 2021. "How dare you call her a pig, I know several pigs who would be upset if they knew"*. *Journal of Language Aggression and Conflict*, 9(1):47–75.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5988–6008. PMLR. ISSN: 2640-3498.

Christian Chukwuebuka Ezeibe and Okey Marcellus Ikeanyibe. 2017. Ethnic Politics, Hate Speech, and Access to Political Power in Nigeria. *Africa Today*, 63(4):65.

Farshid Faal, Jia Yuan Yu, and Ketra A. Schmitt. 2021. Domain adaptation multi-task deep neural network for mitigating unintended bias in toxic language detection. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence, ICAART 2021, Volume 2, Online Streaming, February 4-6, 2021*, pages 932–940. SCITEPRESS.

N. Farrington, L. Hall, D. Kilvington, J. Price, and A. Saeed. 2014. *Sport, racism and social media*. Routledge.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA Evaluation of NLP and Speech Tools for Italian Proceedings of the Final Workshop 12-13 December 2018, Naples*. Accademia University Press.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Tamara Fuchs and Fabian Schäfer. 2020. Normalizing misogyny: hate speech and verbal abuse of female politicians on Japanese Twitter. *Japan Forum*, pages 1–27.

Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities*, 22(3):324–341. Publisher: Routledge _eprint: https://doi.org/10.1080/13504630.2015.1128810.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and Detecting Abusive Language Across Domains and Languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Genevieve Gorrell, Tracie Farrell, and Kalina Bontcheva. 2020. Mp twitter abuse in the age of covid-19: White paper.

Genevieve Gorrell, Mark Greenwood, Ian Roberts, Diana Maynard, and Kalina Bontcheva. 2018. Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians and twaddle: Trends in online abuse towards UK politicians. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Mark A. Greenwood, Mehmet E. Bakir, Genevieve Gorrell, Xingyi Song, Ian Roberts, and Kalina Bontcheva. 2019. Online abuse of uk mps from 2015 to 2019: Working paper.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Kilem L. Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Hexiang Hu, Ozan Sener, Fei Sha, and Vladlen Koltun. 2020. Drinking from a firehose: Continual learning with web-scale natural language. Version: 2.

Sean Ingle. 2021. Sports bodies to boycott social media for bank holiday weekend over abuse. *The Guardian*.

David V. James, Frank R. Farnham, Seema Sukhwal, Katherine Jones, Josephine Carlisle, and Sara Henley. 2016. Aggressive/intrusive behaviours, harassment and stalking of members of the United Kingdom parliament: a prevalence study and cross-national comparison. *The Journal of Forensic Psychiatry & Psychology*, 27(2):177–197.

Adam Joinson, Katelyn Y. A. McKenna, Tom Postmes, and Ulf-Dietrich Reips. 2009. *Oxford Handbook of Internet Psychology*. Oxford University Press.

Daniel Kilvington and John Price. 2019. Tackling Social Media Abuse? Critically Assessing English Football's Response to Online Racism. *Communication & Sport*, 7(1):64–79. Publisher: SAGE PublicationssSage CA: Los Angeles, CA.

Anthony King. 2003. *The European Ritual: Football in the New Europe*. Ashgate Publishing Ltd.

Colin King. 2004. *Offside racism: Playing the white man*. Routledge.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022a. Handling and Presenting Harmful Text in NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Kirk, Bertie Vidgen, Paul Röttger, Tristan Thrush, and Scott Hale. 2022b. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368.

Hannah Rose Kirk, Bertie Vidgen, and Scott A. Hale. 2022c. Is More Data Better? Using Transformers-Based Active Learning for Efficient and Effective Detection of Abusive Language. In *Proceedings of the 3rd workshop on Threat, Aggression and Cyberbullying (COLING 2022)*. Association for Computational Linguistics.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo, and Yang Liu. 2022. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5454. Association for Computational Linguistics.

Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving Generalization of Hate Speech Detection Systems to Novel Target Groups via Domain Adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Lucy Manning and Phillip Kemp. 2019. MPs describe threats, abuse and safety fears. *BBC News*.

J. Reid Meloy, Lorraine Sheridan, and Jens Hoffmann, editors. 2008. *Stalking, Threatening, and Attacking Public Figures*. Oxford University Press.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Paul E. Mullen, David V. James, J. Reid Meloy, Michele T. Pathé, Frank R. Farnham, Lulu Preston, Brian Darnley, and Jeremy Berman. 2009. The fixated and the pursuit of public figures. *Journal of Forensic Psychiatry & Psychology*, 20(1):33–47. Publisher: Routledge.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283.

UK Parliament. 2023. Women in Parliament Today. Technical report, UK Parliament.

Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong Learning of Hate Speech Classification on Social Media. ArXiv:2106.02821 [cs].

Ludovic Rheault, Erica Rayment, and Andreea Musulan. 2019. Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, 6(1).

Ian Rowe. 2015. Civility 2.0: a comparative analysis of incivility in online political discussion. *Information, Communication & Society*, 18(2):121–138.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. 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.

Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. 2021. "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In *Proceedings of the Web Conference 2021*, pages 1110–1121. ACM.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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.

Stuart W. Shulman. 2009. The Case Against Mass E-mails: Perverse Incentives and Low Quality Public Participation in U.S. Federal Rulemaking. *Policy & Internet*, 1(1):22–52.

Ekaterina Stambolieva. 2017. Methodology : Detecting Online Abuse against Women MPs on Twitter. Technical Report Amnesty International, Amnesty International.

John Suler. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior*, 7(3):321–326.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. *Online harassment*, pages 29–55.

Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Parnet. 2016. A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates. *Journal of Communication*, 66(6):1007–1031.

Bertie Vidgen, Yi-Ling Chung, Pica Johansson, Hannah Rose Kirk, Angus Williams, Scott A. Hale, Helen Zerlina Margetts, Paul Röttger, and Laila Sprejer. 2022. Tracking Abuse on Twitter Against Football Players in the 2021 – 22 Premier League Season.

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

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.

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

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. 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.

Stephen Ward and Liam McLoughlin. 2020. Turds, traitors and tossers: the abuse of UK MPs via Twitter. *The Journal of Legislative Studies*, 26(1):47–73.

Andy Williamson. 2009. The Effect of Digital Media on MPs' Communication with Constituents. *Parliamentary Affairs*, 62(3):514–527.

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.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. 2023. Lahm: Large annotated dataset for multi-domain and multilingual hate speech identification. *arXiv preprint arXiv:2304.00913*.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Lanqin Yuan and Marian-Andrei Rizoiu. 2022. Detect hate speech in unseen domains using multi-task learning: A case study of political public figures. *arXiv preprint arXiv:2208.10598*.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

A Data Annotation

We used two different sets of annotators across the two domains, as we annotated the sets sequentially. Initial annotation rounds revealed high rates of annotator disagreement, with a large number of entries requiring expert annotation as a result. We use the same label schema for all domain and demographic pairs but use specific example tweets in the guidelines. We only employ annotators who pass a test of gold questions. Annotators were informed prior to accepting the task that the data would be used to train machine learning models as part of a research paper.

We employed 3,375 crowdworkers for male footballers and 3,513 for female footballers. Crowdworkers were paid \$0.20 per annotation, earning \$11.30/hour on average. Each entry was annotated by 3 crowdworkers, with an additional two annotations required if no majority agreement ($\frac{2}{3}$) was reached, then sent for expert annotation if still no majority agreement ($\frac{3}{5}$) was reached. The average annotator agreement per entry was 68%, and the Cohen's kappa was 0.50.

For the MP datasets, we employed 23 high-quality annotators from a Trust & Safety organisation. Annotators were paid \$0.33 per annotation, earning \$16.80/hour on average. Each entry received 3 annotations, then sent for expert annotation if no majority agreement was reached ($\frac{2}{3}$). The average entry-wise agreement was 82% and the Cohen's kappa was 0.67.

An example of instructions given to annotators is displayed in Figure 6. Fictional examples of tweet stances across domain-demographic pairs are visible in Figure 7. Due to the potentially harmful nature of the task, annotators were encouraged to regularly take breaks, and to contact their line manager in event of any problems or concerns. Annotator pay was above US minimum hourly wage on average.

B Data Statement

To document the generation and provenance of our dataset, we provide a data statement below (Bender and Friedman, 2018).

Curation Rationale The purpose of the DoDo dataset is to train, evaluate, and refine language models for classification tasks related to understanding online conversations directed at footballers and MPs.

Language Variety Due to the UK-centric domains this dataset concerns (men's and women's UK football leagues, and UK MPs), all tweets are in English.

Speaker Demographics All entries are collected from Twitter and therefore generally represent the demographics of the platform. The sample is skewed towards those engaging in community discussion of the two domains on the platform (sports and politics).

Annotator Demographics The two domains used differing annotator pools. For the MPs data, we made use of a company offering annotation services that recruited 23 annotators to work for 5 weeks in early 2023. The annotators were screened from an initial pool of 36 annotators who took a test consisting of 36 difficult gold-standard questions (containing examples of all four class labels). The annotators had constant access to both a core team member from the service provider and from the core research team.

Fifteen annotators self-identified as women, and eight as men. The annotators were sent an optional survey to provide further information on their demographics. Out of 23 annotators, 21 responded to the survey. By age, 12 annotators were between 18-29 years old, eight were between 30-39 years old, and one was over 50 years old. In terms of completed education level, three annotators had high school degrees, eight annotators had undergraduate degrees, six annotators had postgraduate taught degrees, and four annotators had postgraduate research degrees. The majority of annotators were British (17), and other nationalities included Indian, Swedish, and United States. Twelve annotators identified as White, with one identifying as White Other and one identifying as White Arab. Other ethnicities included Black Caribbean (1), Indian (1), Indian British Asian (1), and Jewish (1). Most annotators identified as heterosexual (14), with other annotators identifying as bisexual (3), gay (1), and pansexual (1). Two chose not to disclose their sexuality. The majority stated that English was their native language (16), and four stated they were not native but fluent in the language. One chose not to disclose whether they were native English speakers or not. The majority of annotators disclosed that they spend 1-2 hours per day on social media (12). Four annotators stated that they spent, on average, less than 1 hour on social media per day (but more than 10 minutes), and five stated they spend more than 2 hours per day on social media. Some of the annotators reported having themselves been targeted by online abuse (9), with 11 reporting 'never' and one preferring not to say.

The datasets for footballers were annotated separately using a crowdsourcing platform. Due to this, we have significantly less detail on the demographics of the users. The fb-m dataset was annotated by 3,375 crowdworkers from 41 countries. The fb-w dataset was annotated by 3,513 crowdworkers from 48 countries. The annotators for both datasets were primarily from Venezuela (56% and 64% respectively) and the United States (29% and 18% respectively).

Speech Situation The data consists of short-form written textual entries from social media (Twitter). These were presented and interpreted in isolation for labelling, i.e., not in a comment thread and without user/network or any additional information.

Text Characteristics The genre of texts is a mix of abusive, critical, positive, and neutral social media entries (tweets).

C Data Collection, Processing, and Sampling

We chose to collect data on members of parliament and footballers: two types of well known public figure that both receive considerable amounts of online abuse but which operate in very different domains. These two domains also serve as useful bases because they have demographic diversity (in particular, they have both male and female participants, with gender being a well known source of difference in terms of abuse being received).

We collect all tweets mentioning a public figure account, keeping only those that either directly reply to tweets written by public figures, or directly mention a public figure account without replying or referencing another tweet. We term these tweets *audience contact*. From the audience contact tweets, we only consider tweets that contain some English text content aside from mentions and URLs. Where the Twitter API Filtered Stream endpoint did not return sufficient data for constructing an unlabelled pool, as was the case for female footballers, we made use of the Twitter API Full Archive Search endpoint to collect historic tweets. Table 5 contains information on the unlabelled pools.

For each domain-demographic pair, starting with the unlabelled pool, we randomly sample (and remove) 3,000 entries for the test set and 1,000 entries for the validation set. We then randomly sample (and remove) 1,500 entries for training and concatenate these with a further 1,500 entries containing a keyword from a list of 731 abusive and hateful keywords (750 entries with at least one profanity keyword and 750 with at least one identity keyword), such that each training set has $3,000$ entries total. The list of keywords is compiled from Davidson et al. (2017); ElSherief et al. (2018); Vidgen et al. (2021b); Kirk et al. (2022b) and is available at [REDACTED]. Each training set has 3,000 entries in total. Table 6 describes the counts of Tweets by stance for each sampling strategy used in the construction of datasets.

We replace all user mentions within tweets with tokens relating to the domain of the public figure mentioned before tweet annotation and use in training models. This does not completely anonymise tweets, as it does not account for other uses of names in tweet text.

D Additional Results

D.1 Where Unseen Performance Exceeds Seen Performance

There are three cases where performance on unseen `dodos` exceeds performance on seen `dodos` in both full and fixed budget scenarios, visible in Figure 1. All three cases include `fb-m` in the training data, suggesting that the `fb-m` test set is more difficult that other `dodos`, or potentially that the `fb-m` training split is significantly different to the test split - further investigation is needed to fully understand this dynamic.

D.2 Error Analysis

Our error analysis is based on each fixed-budget single `dodo` model (i.e. `dodo1` experiments), evaluated on seen portions of the test set. We also analyse errors made by the fixed budget generalist model (i.e. `dodo4`), and shared errors made by all fixed budget condition models. We choose fixed budget models to ensure all models have seen the same total amount of training data. We present confusion matrices for all experiments in Fig. 8.

The fb-m model performed best on positive tweets (F1 = 0.86), and worst on critical tweets (F1 = 0.52). These results broadly hold for the fb-w model, which performed best on positive tweets (F1 = 0.91) and less well on abusive (F1 = 0.57) and critical (F1 = 0.52) tweets. The mp-m model performed best on critical tweets (F1 = 0.77), and worst on positive and neutral tweets (F1 = 0.69). As with footballers, these results broadly hold for the mp-w model, which performed best on critical tweets (F1 = 0.74), and less well on neutral (F1 = 0.66) and abusive tweets (F1 = 0.63).

These results partly reflect class imbalance (the FBs data is heavily skewed towards positive tweets, the MPs data towards critical tweets), as well as some inherent similarity between classes which border one another i.e., positive vs. neutral, neutral vs. critical, and critical vs. abusive. Recurring errors reveal several tweet types that are challenging to classify: tweets that tweets that (i) contain a mixture of both positive and critical language; (ii) use positive or sarcastic language to mock; (iii) rely on emoji to convey abuse; (iv) contain niche insults; or (v) short, ambiguous tweets that lack context.

D.3 Expanded Evaluation

Here we provide expanded reference tables and figures on the results described in Section 4.

The per-class macro F1 score of each dodo1 model and the two dodo4 models evaluated on seen dodos are visible in Table 7, revealing relatively low performance on the critical and abusive classes for models trained on the two footballer datasets compared to the positive and neutral classes. For models trained on the MPs datasets, we see much less variation in per class performance.

We also present a set of confusion matrices in Figure 8 for the specialist (dodo1), fixed budget generalist (dodo4, train size = 3,000), and full budget generalist (dodo4, train size = 12,000) models based on deBERT, evaluated on each evaluation set and the total evaluation set.

Finally, we give a reference table of maximum Macro-F1 scores achieved by all baselines across all evaluation sets (Table 8).

dodo	**Per-class F1 Scores**			
	Positive	*Neutral*	*Critical*	*Abusive*
fb_m	0.86	0.66	0.52	0.58
fb_w	0.94	0.81	0.57	0.62
mp_m	0.69	0.69	0.77	0.70
mp_w	0.72	0.66	0.74	0.63
All (fixed)	0.87	0.67	0.71	0.61
All (raw)	0.89	0.71	0.73	0.66

Table 7: Per-class F1-scores for dodo1 and dodo4 baselines on seen evaluation sets.

Overview

Content Warning: This task contains examples of hateful and abusive tweets. Please take frequent breaks during annotation, and contact your line manager for support.

This is a task annotating tweets relating to and discussing football (soccer) and politicians (MPs). The goal is to identify the sentiment of language used in the tweets (the options are: abusive, critical, neutral or positive).

Apply the coding guidelines dispassionately and try to mitigate any personal biases you hold.

Only tweets in English should be annotated. If it is clearly NOT in English then flag this. Tweets with one-off non-english words still counts as Yes.

Task

Select one option which best describes the tone of language in the tweet: *abusive, critical, neutral or positive*. Definitions of these options can be found below. When you consider the stance/sentiment, make sure to take into account all signals of a tweet's tone such as capitalization, punctuation and emoji. If the tweet has two parts with different stances, pick the stance which dominates the tone.

Stance	Definition
Abusive	**Select IF: the tweet threatens, insults, derogates (e.g. hateful use of slurs, negative use of stereotypes), dehumanises (e.g. compares individuals to insects, animals or trash), mocks or belittles an individual or their identity.** **Note on distinguishing between Abusive and Critical:** Criticism, discussion and incivility are not the same as abuse. If the tweet does not use aggressive language, or if it makes a substantive criticism of an individual or group of individuals, it should be marked as 'Critical'. For example, "*And let's not forget that idiot leader we got [USER]. This has been going on for too long.*" should be marked as Critical, not Abusive, because the dominating tone of the tweet is critical even though the person has been called an 'idiot'.
Critical	**Select IF: the tweet makes a substantive criticism of an individual or small groups of individuals.** This could include critique of their behavior or their actions. Criticism is not a form of 'soft abuse'. For a tweet to be legitimate criticism, it must not use slurs or aggressive and insulting language. **Note on Abusive/Critical**: The language used can be emotive and still be critical, for example: "*How the fucking hell is that not a red card. Absolutely sickening challenge from [PLAYER]*" . However, if it becomes aggressive, demeaning or insulting, then the tweet should be marked as 'Abusive'. Criticism of an individual purely on the basis of their identity, should be marked as 'Abusive', for example claiming a player is bad because of their race.
Neutral	**Select IF: the tweet makes no emotional or sentimental comment towards a person or an identity.** Neutral statements could include unemotive factual statements or descriptions of events. **Note on Lacking information:** If the tweet has very little context to decide the stance, mark it as neutral e.g. if it only uses one emoji with no clear context.
Positive	**Select IF: the tweet supports, praises or encourages a person or identity.** It can include support, respect or encouragement of a particular skill, behavior, achievement or success, or positive views towards diversity and representation of identities like race and sexuality.

Figure 6: Instructions given to annotators.

	Positive	Neutral	Critical	Abusive
Footballers Men	[PLAYER] [USER] CR7 GOAT!!	[PLAYER] puts [CLUB] 1-0 up against [CLUB] [URL] #goal	It wouldn't be so hard to watch [CLUB] if [PLAYER] didn't bottle it every time #coys	[PLAYER] get out of my club shithead
Footballers Women	Love you you absolute beast [PLAYER]	[PLAYER] You'll get used to the cold eventually!	[PLAYER] who keeps telling you you should be taking pens, it's painful to watch	[PLAYER] fuck off
MPs Men	[MP] great speech sir	Does anyone else think [MP] and [MP] look strangely similar? #doppelganger	[MP] Why should anyone believe you when everything you say gets proven to be a lie?	[MP] Who the fuck voted you in scumbag #corrupt
MPs Women	[MP] you're one of the good ones	[MP] [USER] Take a look at the report shared by [MP], pretty stark numbers	[MP] good one, talk about dignity when you and your colleagues spent it all on filling your own pockets...	[MP] Turns out this bitch is blind as well as stupid

Figure 7: Fictional example tweets for each class label, loosely based on topics and sentiment of content in the dataset. Entries from the dataset are presented to annotators as shown, with special tokens to represent tagged mentions of public figures, accounts representing affiliations (e.g., football clubs), and other users. Examples are fictional as the dataset will not be released.

Domain	Demographic	Pool Size	Collection Dates		Collection Method	
			Start	*End*	*Streaming*	*Search*
Footballers	Men	1,008,399	12/08/2021	02/02/2022	✓	
	Women	226,689	13/08/2021	28/11/2022	✓	✓
MPs	Men	1,000,000	13/01/2022	19/09/2022	✓	
	Women	1,000,000	13/01/2022	19/09/2022	✓	

Table 5: Dates and pool sizes for each domain-demographic pair.

Split	dodo	**Sampling Strategy**											
		Random				Profanity Keywords				Identity Keywords			
		Abusive	*Critical*	*Neutral*	*Positive*	*Abusive*	*Critical*	*Neutral*	*Positive*	*Abusive*	*Critical*	*Neutral*	*Positive*
Train	fb_m	45	172	531	752	290	224	52	184	532	79	64	75
	fb_w	18	63	432	987	346	190	211	467	117	29	76	64
	mp_m	212	725	471	92	372	311	57	10	423	247	77	3
	mp_w	153	746	477	124	349	322	67	12	368	285	84	13
Test	fb_m	103	377	811	1709	0	0	0	0	0	0	0	0
	fb_w	43	89	767	2101	0	0	0	0	0	0	0	0
	mp_m	392	1467	985	156	0	0	0	0	0	0	0	0
	mp_w	373	1471	927	229	0	0	0	0	0	0	0	0
Validation	fb_m	33	93	335	539	0	0	0	0	0	0	0	0
	fb_w	14	45	267	674	0	0	0	0	0	0	0	0
	mp_m	140	484	332	44	0	0	0	0	0	0	0	0
	mp_w	135	459	337	69	0	0	0	0	0	0	0	0
Total	**fb_m**	**181**	**642**	**1677**	**3000**	**290**	**224**	**52**	**184**	**532**	**79**	**64**	**75**
	fb_w	**75**	**197**	**1466**	**3762**	**346**	**190**	**211**	**467**	**117**	**29**	**76**	**64**
	mp_m	**744**	**2676**	**1788**	**292**	**372**	**311**	**57**	**10**	**423**	**247**	**77**	**3**
	mp_w	**661**	**2676**	**1741**	**422**	**349**	**322**	**67**	**12**	**368**	**285**	**84**	**13**

Table 6: Tweet counts for dodo splits across sampling strategy and stance.

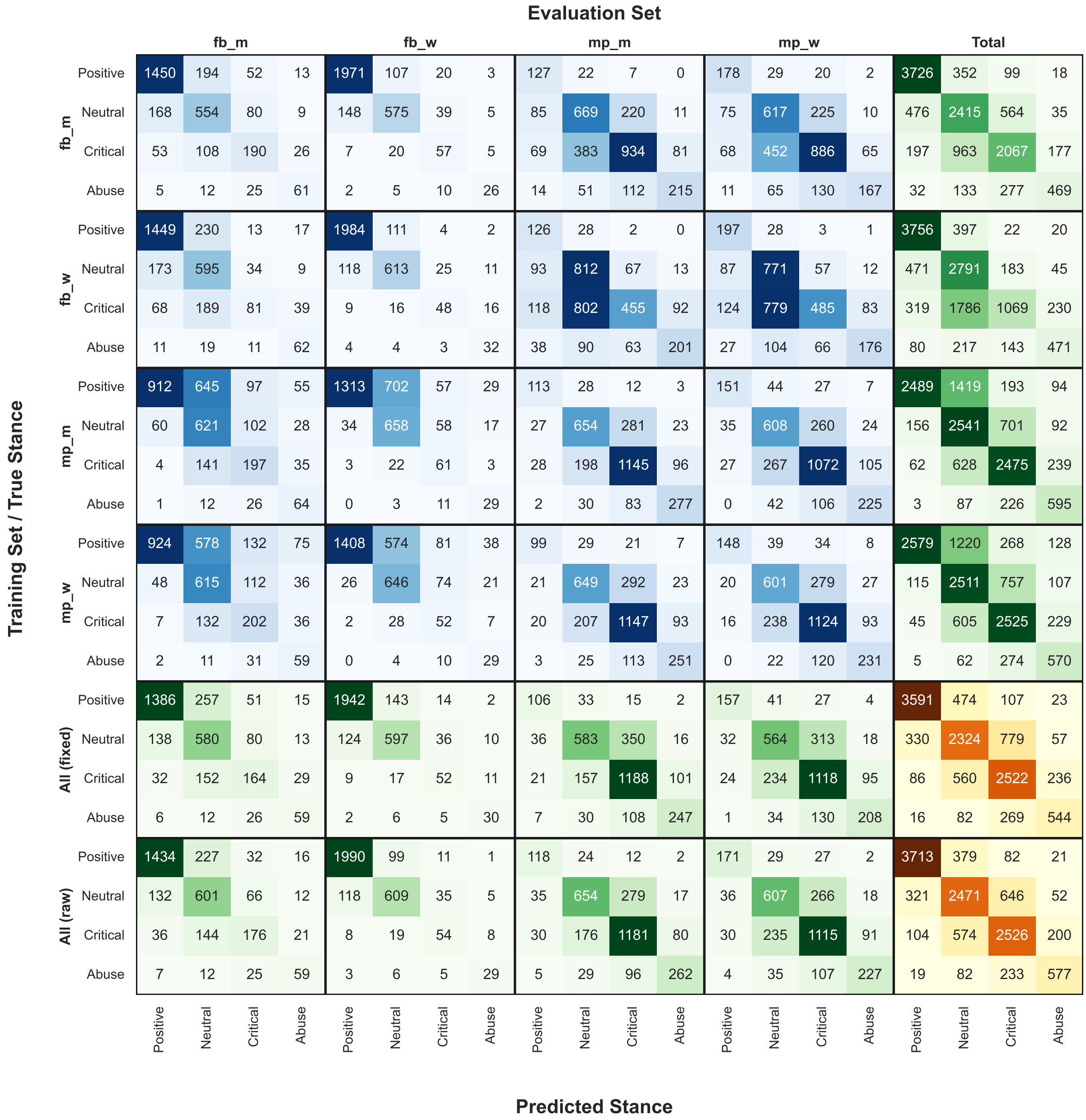

Figure 8: Grid of confusion matrices across chosen baselines, using soft voting across random seeds.

	Train On						Test On				
	fb-m	fb-w	mp-m	mp-w	model	budget	total	fb-m	fb-w	mp-m	mp-w
dodo1	✓				deBERT	fixed = full	0.688	0.656	0.719	0.633	0.609
	✓				diBERT	fixed = full	0.600	0.580	0.589	0.518	0.522
		✓			deBERT	fixed = full	0.628	0.586	0.676	0.539	0.545
		✓			diBERT	fixed = full	0.508	0.476	0.615	0.415	0.413
			✓		deBERT	fixed = full	0.665	0.536	0.576	0.71	0.665
			✓		diBERT	fixed = full	0.571	0.438	0.437	0.619	0.587
				✓	deBERT	fixed = full	0.675	0.549	0.578	0.681	0.683
				✓	diBERT	fixed = full	0.584	0.449	0.446	0.592	0.605
dodo2	✓	✓			deBERT	fixed	0.668	0.637	**0.790***	0.588	0.579
	✓	✓				full	0.668	0.639	0.709	0.596	0.594
	✓	✓			diBERT	fixed	0.577	0.557	0.593	0.494	0.501
	✓	✓				full	0.611	0.586	0.61	0.521	0.519
	✓		✓		deBERT	fixed	0.713	0.634	0.722	0.686	0.657
	✓		✓			full	0.724	0.659	0.705	0.704	0.669
	✓		✓		diBERT	fixed	0.652	0.568	0.588	0.602	0.594
	✓		✓			full	0.671	0.598	0.608	0.613	0.61
	✓			✓	deBERT	fixed	0.715	0.646	0.665	0.691	0.671
	✓			✓		full	0.724	0.658	0.69	0.694	0.681
	✓			✓	diBERT	fixed	0.647	0.564	0.587	0.58	0.595
	✓			✓		full	0.665	0.59	0.594	0.611	0.613
		✓	✓		deBERT	fixed	0.703	0.606	0.694	0.671	0.646
		✓	✓			full	0.721	0.608	0.699	0.71	0.669
		✓	✓		diBERT	fixed	0.647	0.494	0.615	0.581	0.575
		✓	✓			full	0.639	0.496	0.575	0.604	0.589
		✓		✓	deBERT	fixed	0.708	0.604	0.679	0.66	0.667
		✓		✓		full	0.722	0.612	0.687	0.695	0.684
		✓		✓	diBERT	fixed	0.629	0.512	0.569	0.567	0.571
		✓		✓		full	0.638	0.511	0.575	0.591	0.611
			✓	✓	deBERT	fixed	0.664	0.533	0.556	0.672	0.683
			✓	✓		full	0.683	0.559	0.575	0.692	0.687
			✓	✓	diBERT	fixed	0.574	0.454	0.416	0.609	0.598
			✓	✓		full	0.624	0.492	0.499	0.634	0.63
dodo3	✓	✓	✓		deBERT	fixed	0.71	0.629	0.737	0.67	0.649
	✓	✓	✓			full	0.721	0.623	0.736	0.701	0.664
	✓	✓	✓		diBERT	fixed	0.636	0.552	0.598	0.576	0.565
	✓	✓	✓			full	0.659	0.577	0.611	0.616	0.591
	✓	✓		✓	deBERT	fixed	0.698	0.614	0.723	0.635	0.636
	✓	✓		✓		full	0.734	0.648	0.726	0.694	0.682
	✓	✓		✓	diBERT	fixed	0.625	0.534	0.576	0.553	0.55
	✓	✓		✓		full	0.672	0.576	0.634	0.591	0.605
	✓		✓	✓	deBERT	fixed	0.713	0.626	0.671	0.685	0.673
	✓		✓	✓		full	**0.736***	**0.664***	0.706	0.712	**0.692***
	✓		✓	✓	diBERT	fixed	0.648	0.557	0.587	0.602	0.609
	✓		✓	✓		full	0.674	0.583	0.593	0.633	0.626
		✓	✓	✓	deBERT	fixed	0.695	0.585	0.663	0.653	0.658
		✓	✓	✓		full	0.724	0.591	0.694	**0.716***	**0.692***
		✓	✓	✓	diBERT	fixed	0.642	0.488	0.569	0.592	0.602
		✓	✓	✓		full	0.663	0.516	0.586	0.614	0.618
dodo4	✓	✓	✓	✓	deBERT	fixed	0.707	0.64	0.703	0.663	0.654
	✓	✓	✓	✓		full	0.728	0.634	0.713	0.709	0.684
	✓	✓	✓	✓	diBERT	fixed	0.644	0.533	0.591	0.58	0.579
	✓	✓	✓	✓		full	0.685	0.589	0.639	0.633	0.633

Table 8: Macro-F1 score for all sets of baseline models (maximum value across three seeds). Best Macro-F1 per test set (total and each of the four dodo splits) is bold and starred. Colour-coded according to increasing Macro-F1 Score.