Studying Generalisability Across Abusive Language Detection Datasets

Steve Durairaj Swamy and **Anupam Jamatia**
Department of Computer Science
National Institute of Technology
Agartala, India
{steve050798,anupamjamatia}@gmail.com

Björn Gambäck*
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
gamback@ntnu.no

Abstract

Work on Abusive Language Detection has tackled a wide range of subtasks and domains. As a result of this, there exists a great deal of redundancy and non-generalisability between datasets. Through experiments on cross-dataset training and testing, the paper reveals that the preconceived notion of including more non-abusive samples in a dataset (to emulate reality) may have a detrimental effect on the generalisability of a model trained on that data. Hence a hierarchical annotation model is utilised here to reveal redundancies in existing datasets and to help reduce redundancy in future efforts.

1 Introduction

With the growth of the internet and the increasingly smaller barrier to entry, social media have become viable platforms for people to make their views known. These easily accessible fora for discourse have given a voice to many minorities and individuals to share their stories. The caveat, however, is that these platforms can be misused to spread hate and harass other individuals, which has given birth to terms such as cyberbullying and trolling. Online harassment has been a point of criticism levied against social media giants such as Facebook and Twitter, who have come under increased pressure to address this misuse. To this end, they have ensured that their community guidelines explicitly ban the usage of profanity/hate speech to harass and bully individuals.

The detection of Online Abuse has proven to be a layered and complex issue. For example, profanity is often treated as a sign of hate speech or offensive language, but profanity can also be used in a wide variety of expressive ways to convey informality, humour, and emphasis. This usage of profanity outside of abuse/insults, coupled with implicit insults that may not contain any profanity, makes the task of classifying abuse online a balancing act of sorts, forming the crux of what makes this task hard to tackle: stricter guidelines may hamper a well-meaning individual's freedom of speech, while more lenient guidelines may empower those who exploit them.

As it stands, the intricacies of free speech do not translate well to machine understanding. This has led to the continued use of human moderators in the abusive language detection space. Content is flagged by users, reviewed by a human and removed if it violates the platform's community guidelines. The main problem with this system is the sheer volume of content to be reviewed, giving human moderators very little time to arrive at a decision. Another issue that was highlighted by Roberts (2019) is the impact that reviewing online abuse can have on a worker's mental well-being. These issues have led to many social media giants, such as Facebook, to seek machine learning-based solutions — to replace or supplement the current human moderator system.

Automatic detection of abusive language online can be seen as a union of the plethora of subtasks that have been tackled: Cyberbullying, Hate Speech (also further constrained as racism, sexism, and harassment of particular minorities), Trolling, etc. Research in the field tends to focus on one of the particular subtasks. It has been argued by some (Schmidt and Wiegland, 2017; Waseem et al., 2017b) that due to this phenomenon where works tackle restricted subsets of abusive language, it has become difficult to make judgements about whether the features being used can perform well in other subtasks of abusive language detection — as they are often only evaluated on a single dataset, specific to one domain and subtask, and annotated in a specific way.

*Also at: RISE SICS, Kista, Sweden.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 940–950
Hong Kong, China, November 3-4, 2019.

Waseem et al. (2017b) proposed that there exists an overlap between these subtasks and subsequently proposed a typology that emphasises identifying the target of abuse and whether the abuse is implicit or explicit. Their typology could potentially be applied to all stages of system development, from data collection to the final model building. This, they hoped, would help to synthesise the different subtasks. This idea was expanded upon in the Offensive Language Identification Dataset (OLID; Zampieri et al. 2019a) to a hierarchical, three-level annotation model.

After further discussing related research in the next section, this work looks at various publicly available datasets in the field (Section 3), and performs both in-domain (Section 4) and cross-dataset training and testing to observe whether models trained on one dataset generalise well when tested against other datasets (Section 5). It also makes some qualitative assessments on why models trained on specific datasets generalise better than others. Additionally, the OLID dataset based on the typology by Waseem et al. (2017b) is used to observe whether the hierarchical annotation model is sufficient to synthesise the various subtasks of abusive language detection. To this end, experiments were run using BERT, Bidirectional Encoder Representations from Transformers (Devlin et al., 2018), to compare its performance to other popular models that have been used for abusive language detection (Section 6).

2 Previous Work

Abusive language detection has served as an umbrella term for a wide variety of subtasks. Research in the field has typically focused on a particular subtask: Hate Speech (Davidson et al., 2017; Founta et al., 2018; Gao and Huang, 2017; Golbeck et al., 2017), Sexism/Racism (Waseem and Hovy, 2016), Cyberbullying (Xu et al., 2012; Dadvar et al., 2013), Trolling and Aggression (Kumar et al., 2018a), and so on. Datasets for these tasks have been collected from various social media platforms, such as Twitter (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018; Burnap and Williams, 2015; Golbeck et al., 2017), Facebook (Kumar et al., 2018a), Instagram (Hosseinmardi et al., 2015; Zhong et al., 2016), Yahoo! (Nobata et al., 2016; Djuric et al., 2015; Warner and Hirschberg, 2012), YouTube (Dinakar et al., 2011), and Wikipedia (Wulczyn et al., 2017), with annotation typically carried out on crowdsourcing platforms such as CrowdFlower (Figure Eight)[1] and Amazon Mechanical Turk.[2]

All these datasets represent multi-class classification problems, with the exception of the Kaggle's Toxic Comment Classification challenge,[3] which entails multi-label classification, and OLID (Zampieri et al., 2019a) used in the SemEval-2019 'OffensEval' shared task (Zampieri et al., 2019b), which builds on a hierarchical annotation model (Hierarchy of Multi-Class Classifiers).

Choice of features has been the crucial difference between the various approaches to abusive language detection. For the most part, word-level n-grams have been highly predictive, with other linguistic features such as part-of-speech tags (Xu et al., 2012; Davidson et al., 2017) and sentiment score (Van Hee et al., 2015; Davidson et al., 2017) providing slight improvements. Due to their ability to perform better in an online setting where spelling errors and adversarial behaviour are commonplace, character-level features have been endorsed (Mehdad and Tetreault, 2016), and also shown to often be superior to word-level information for this task (Meyer and Gambäck, 2019). Metadata about users have also been used as features: Waseem and Hovy (2016) claim gender information leads to improved performance, while Unsvåg and Gambäck (2018) report user-network data to be more important. Schmidt and Wiegland (2017) provides a comprehensive overview of many of the features used and their efficacy.

In terms of models, popular classical classification approaches include Logistic Regression and LSVM (Linear Support Vector Machines). Deep Neural networks such as Convolutional Neural Networks, CNN (Zhang et al., 2018; Gambäck and Sikdar, 2017) and variations of Recurrent Neural Networks, RNN (Pitsilis et al., 2018; Gao and Huang, 2017) have seen widespread success, regularly obtaining state-of-the-art results on various datasets. Lee et al. (2018) used the Founta et al. (2018) dataset to conduct a comparative study of the performance of many popular models. In the 'OffensEval' shared task (Zampieri et al., 2019b), the use of contextual embeddings such as BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018) exhibited the best results.

[1] figure-eight.com
[2] mturk.com
[3] bit.ly/2HNfLaB

Generalisability of a model has also come under considerable scrutiny. Works such as Karan and Šnajder (2018) and Gröndahl et al. (2018) have shown that models trained on one dataset tend to perform well only when tested on the same dataset. Additionally, Gröndahl et al. (2018) showed how adversarial methods such as typos and word changes could bypass existing state-of-the-art abusive language detection systems. They also observed unimpressive results when using ULMFiT (Howard and Ruder, 2018) for abusive language detection, but argued that model architecture is less important than the type of data and the annotation scheme.

Karan and Šnajder (2018) experimented with cross-domain training and testing, and opted to use the same model (LSVM) with minimal features and to preprocess in favour of interpretability. They also reported positive improvements using Frustratingly Easy Domain Adaptation (FEDA; Daumé III, 2007) to augment smaller datasets with larger ones. Fortuna et al. (2018) concurred, stating that although models perform better on the data they are trained on, slightly improved performance can be obtained when adding more training data from other social media. Similarly, Waseem et al. (2018) attempted to address the problem of differences between datasets by building a robust multi-task learning model, which improves upon single-task performance by using auxiliary samples from select datasets. Their work revealed that such models could be competitive with the state-of-the-art single-task models with the additional benefit of allowing prediction on other datasets as well. This helps in negating hidden biases within datasets and promoting generalisability.

3 Datasets

The experiments in the next section will be based on four different datasets, annotated for hate speech and/or offensive language, as described below. The social media platform of choice, Twitter, was selected due to the availability of a multitude of easy to access datasets. The datasets are all in English and from Twitter, and largely chosen based on popularity and availability.

The first two datasets, from Waseem and Hovy (2016) and from Davidson et al. (2017) were chosen due to their widespread use as benchmarks for models. The third, from Founta et al. (2018) was selected because of its large size, while the fourth (Zampieri et al., 2019a) was included since it is using the contemporary hierarchical model. Some other large datasets were discarded since they are either not from Twitter (such as the Kaggle Toxicity classification of Wikipedia comments, Wulczyn et al., 2017) or not easily or openly available (e.g., Silva et al., 2016; Golbeck et al., 2017).

3.1 The Waseem and Hovy Dataset

In their work on the disambiguation of types of hate speech, Waseem and Hovy (2016) released a dataset of 16,914 tweets. They solicited their tweets using a lexicon of hate speech terms, and manually annotated them with three tags: `racism`, `sexism`, and `none`. Waseem and Hovy used an expert outside annotator for reviewing their annotations to mitigate any bias. The database is provided as a set of tweet IDs with tags, but many of the actual tweets have been removed over time, in particular those belonging to the racist class.[4] The first set of rows in Table 1 describes the dataset, including a comparison of the original Waseem and Hovy (2016) dataset to the one available for download using the Twitter API when the present experiments were initiated.

3.2 The Davidson et al. Dataset

Davidson et al. (2017) made publicly available a Twitter dataset with three labels: `hate_speech`, `offensive_language`, and `neither`. Similar to Waseem and Hovy (2016), they used a lexicon of hate speech terms derived from `Hatebase.org` and queried Twitter using these terms to collect potentially hateful tweets. Each tweet was annotated by at least three CrowdFlower workers and the tags were assigned based on the majority decisions. The final dataset available online contains 24,783 tweets. Table 1 provides some statistics of the dataset, which henceforth will be referred to as the Davidson et al. dataset.

Note the very large fraction of abusive tweets in the dataset. A possible explanation for this was given by Waseem et al. (2018), who noted that 2,161 tweets in Davidson et al.'s dataset written in African American Vernacular English had been annotated as offensive or hateful when including the n-word, although the actual usage was to mark group inclusion and informality. While Waseem et al. discuss that these errors were due to the

[4]Note that this discrepancy means that comparisons to work by others on this dataset are not straight-forward.

Dataset	Total	Normal	Hatespeech			Offensive / Abusive					Spam
				racism	sexism						
Waseem and Hovy											
original	16,914	11,559	5,355	1,972	3,383	N/A					N/A
available	11,112	8,185	2,927	17	2,910	N/A					N/A
Davidson et al.	24,783	4,163	1,430			19,190					N/A
Founta et al.	99,996	53,851	4,965			27,150					14,030
							UNT	TIN (targeted)			
								IND	GRP	OTH	
Zampieri et al.	14,100	9,460	N/A			4,640	551	2,507	1,152	430	N/A

Table 1: Overview of the datasets by Davidson et al., Founta et al., Waseem and Hovy, and Zampieri et al.

scarcity of African Americans among the annotators, they could also be attributed to lack of meta-information about the tweet authors: had the annotators known that those tweets were written by African Americans, they would probably have induced that the n-word was not used offensively.

3.3 The Founta et al. Dataset

Founta et al. (2018) released a large Twitter dataset with four labels: `hateful`, `abusive`, `normal`, and `spam`. The main part of their work revolved around a methodology to collect and annotate data over crowdsourcing platforms. They collected tweets from the Live Twitter stream and filtered them using sentiment score (searching for tweets with strong negative polarity) and a lexicon of offensive words from `Hatebase.org` and `noswearing.com/dictionary`.

Table 1 also introduces the Founta et al. dataset, which with a total of $99,996$ tweets is by far the largest in the present study, but also contains a sizable fraction of spam tweets (a category which is not included in the other datasets).

3.4 OLID

The Offensive Language Identification Dataset, OLID (Zampieri et al., 2019a) was used in SemEval-2019 Task 6: 'OffensEval' (Zampieri et al., 2019b). It consists of $14,100$ tweets annotated through a unique hierarchical model whose basic idea was proposed by Waseem et al. (2017b). For the shared task, the data was split into (non-stratified) training and test sets containing $13,240$ and 860 tweets, respectively.

As can be seen in last rows of Table 1, there are three annotation levels in OLID, each of which was directly reflected as a subtask in OffensEval:

A. Whether the tweet can be classified as being offensive (`OFF`) or non-offensive (`NOT`).

B. Tweets labelled as `OFF` are further classified as either `UNT` (untargeted insult/abuse) or `TIN` (targeted insult/abuse).

C. Tweets labelled as `TIN` are sub-divided as `IND` (insults targeted at an individual), `GRP` (insults targeted at a minority group) or `OTH` (insults targeted at an issue or organisation).

4 Preliminary Feature and Model Study

The first set of experiments aimed to test the efficacy of BERT (Devlin et al., 2018) when tackling the Abusive Language Detection task. For this, BERT's performance was compared to three other popular classifiers: Linear SVM, an LSTM (Long Short-Term Memory) Recurrent Neural Network (Hochreiter and Schmidhuber, 1997), and ELMo (Peters et al., 2018). The methodology and models are briefly explained here.

To shed some light on the models themselves rather than the features, no extra surface-level features or linguistic features were utilised in the classification. Also preprocessing was minimal, with lower-casing of tweets being the only standard. However, fine-tuning was carried out on the models' hyper-parameters, such as sequence length, drop out, and class weights. Test and training sets were created for each dataset by performing a stratified split of 20% vs 80%, with the larger part used for training the models. The training sets were further subdivided, keeping 1/8 shares of them as separate validation sets during development and fine-tuning of the hyper-parameters. However, the validation sets were conflated with the training sets for the final results as some of the datasets were already quite small and the models benefited from the extra data. Information on the models themselves are provided below.

Dataset	LSVM		LSTM		ELMo		BERT	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Waseem and Hovy	.8911	.5696	.8498	.5312	.8614	.5394	**.9023**	**.5837**
Davidson et al.	.9014	.7278	.9143	.7419	.8909	.6802	**.9172**	**.7727**
Founta et al.	.8034	.6591	.8161	.6788	.8094	.6732	**.8187**	**.6960**
OLID (Subtask A)	.7610	.7068	.7894	.7479	.7663	.7198	**.8004**	**.7738**

Table 2: Model test tesults (macro-F_1 and accuracy) for all datasets; the best performer is in bold.

4.1 Linear SVM

The Linear SVM (LSVM) was modelled and trained in the `Scikit-learn`[5] library (Pedregosa et al., 2011), utilising a TF-IDF vector representation for the tweets. The classes were artificially balanced and overfitting penalised using L2 regularisation. Interesting hyperparameters included the n-gram range and whether to use character or token n-grams. For example, the Davidson et al. dataset tended to perform better with token n-grams, while the Waseem and Hovy dataset worked better with character n-grams. The inclusion of unigrams was also pivotal to good classifier performance when using token n-grams.

4.2 LSTM Network

The tested Deep Learning Model was built on a fairly simple LSTM architecture using Keras[6] with a TensorFlow[7] back end. The 'Adam' optimiser (Kingma and Ba, 2014) was paired with categorical cross-entropy loss function for model training. Again no statistical or linguistic features were used and the only preprocessing involved lower-casing the tweets. The first layer used a 200 dimensional `GloVe embedding`,[8] pre-trained on 2 billion tweets (Pennington et al., 2014), with embedding weights fixed throughout the training. The Embedding Layer was followed by an LSTM layer of 200 units. The final layer was a dense layer with softmax activation and layer size dependent on the number of classes in the dataset being tested. The most significant hyperparameters were found to be dropout and class weights.

4.3 ELMo

The third model tested used ELMo for feature extraction and was implemented in the TensorFlow hub module[9] with 1024 dimensional ELMo embeddings. This input was passed through an LSTM layer of dimension 256 and then a dense layer with a softmax activation function. The size of the last dense layer was again equal to the number of labels that should be classified. The 'Adam' optimiser and categorical cross-entropy loss function were used during training. ELMo's standalone performance was found to not be as impressive as hoped, with the batch size and usage of dropout significantly affecting classification rates.

4.4 BERT

$\texttt{BERT}_{\texttt{base, uncased}}$ was used as the underlying pre-trained model, in a fine-tuning only approach with no statistical or linguistic features. The model built on the `run_classifier` API provided on the BERT GitHub page[10] and the BERT tokeniser, which simply lower-cases sentences and removes illegal characters. $\texttt{BERT}_{\texttt{base,uncased}}$ trains a total of 110 million parameters, and contains 12 transformer blocks and 12 self-attention heads with hidden layer dimension 768. The most successful parameter settings utilised larger maximum sequence lengths, but smaller batch sizes and lower learning rates. The best models used a learning rate of e^{-5} and batch size 32 with varying maximum sequence lengths between 60 and 70. Other parameters worth mentioning are the number of epochs and the Linear Warm-up Proportion.

4.5 Results

The experimental results are recorded in Table 2, with most improvements and decrements in performance across models being minimal. BERT exhibits the best results for all datasets used in the experiments (with a significance level of 0.05). Surprisingly, ELMo was neither competitive with BERT nor with the GLoVE-embedding LSTM recurrent neural network (when tested with the same statistical significance level).

[5] `scikit-learn.org/stable/`
[6] `github.com/fchollet/keras`
[7] `tensorflow.org/`
[8] `nlp.stanford.edu/projects/glove/`
[9] `tfhub.dev/google/elmo/2`
[10] `github.com/google-research/bert`

Dataset	Positive labels	Negative labels	Positive label fraction
Waseem and Hovy	racism, sexism	neither	26.34%
Davidson et al.	hate_speech, offensive_language	neither	77.43%
Founta et al.	hateful, abusive	spam, none	32.12%
OLID	OFF	NOT	32.91%

Table 3: Cross-dataset experiment, positive and negative label split.

Dataset	Waseem and Hovy		Davidson et al.		Founta et al.		OLID	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Waseem and Hovy	**.9037**	**.8755**	.5626	.5296	.7205	.5824	.6716	.5982
Davidson et al.	.7719	.6928	**.9639**	**.9351**	.9261	.9157	.7514	.6847
Founta et al.	.7278	.6049	.8324	.7559	**.9421**	**.9340**	.7862	.7447
OLID	.7108	.6269	.8247	.7308	.9251	.9162	**.8004**	**.7738**

Table 4: Cross-dataset test results (accuracy and macro-F_1) for all dataset combinations, using the BERT models. Rows show the dataset used to train the model and columns the dataset used for testing.

5 Cross-Dataset Training and Testing

In the second round of experiments, the best models built for individual dataset were used to test generalisability across the other datasets. For all datasets, these were BERT models, but with varying hyper-parameter settings. Karan and Šnajder (2018) used a simpler Linear SVM model for all the datasets for the sake of interpretability, while the aim here, in contrast, was to see how well the best models (that may have learnt some dataset-specific biases) performed on other datasets. This was done to investigate how well state-of-the-art systems perform in a real-life scenario, i.e., when exposed to data from other domains, with the hypothesis that a model trained on one dataset that exhibits comparatively reasonable results on other datasets can be expected to generalise well.

For these experiments, the models were tested on the test set which had been generated for the preliminary model study described in Section 4. As there exists a large number of heterogeneous annotation schemes between datasets, the same approach as Karan and Šnajder (2018) was taken, separating the tags in each dataset according to positive (abusive) and negative (benign) labels. This separation is represented in Table 3, which also gives the percentage of positive samples in each dataset. As can be seen, three of the datasets contain slightly less than 1/3 abusive instances. The Davidson et al. dataset stands out, by containing 3/4 offensive instances. As discussed in Section 3.2, this can probably be attributed to how those tweets were selected and annotated.

The results of cross-dataset testing are presented in Table 4. Considerable performance drops can be observed when going from a large training dataset to a small test set (i.e., Founta et al.'s results when tested on the Waseem and Hovy dataset) and vice versa. This is in line with a similar conclusion by Karan and Šnajder (2018).

It is surprising to see how well a model trained on Founta et al.'s dataset performs when tested on OLID (Zampieri et al., 2019a) and vice versa. However, this can be expected to be the case where there is a good agreement between the datasets, i.e., there is a large amount of similar data shared between them. To this effect, the Founta et al. dataset was searched with terms used by Zampieri et al. when collecting data for OLID, giving around $6,600$ hits. For comparison, OLID gets around $12,200$ hits with the same set of terms.

The most interesting observation is that datasets with larger percentages of positive samples tend to generalise better than datasets with fewer positive samples, in particular when tested against dissimilar datasets. For example, we see that the models trained on the Davidson et al. dataset, which contains a majority of offensive tags, perform well when tested on the Founta et al. dataset, which contains a majority of non-offensive tags. (The differences are all statistically significant when the test set is Waseem and Hovy.) Similar trends were observed by Karan and Šnajder (2018) when employing the Kolhatkar et al. (2018) and TRAC-1 (Kumar et al., 2018a) datasets, that have 62.7% and 56.6% positive samples, respectively, and exhibited better results in cross-dataset testing than datasets with lower positive sample ratios.

	Subtask A		Subtask B		Subtask C	
	BERT	Top	BERT	Top	BERT	Top
F_1	.8168	.8286	.6997	.7545	.6162	.6597
Acc.	.8546	.8628	.9000	.9250	.7136	.7277
Rank	2	1	9	1	6	1

Table 5: BERT test set results (macro-F_1 and accuracy) compared to top OffensEval shared task performers.

6 Synthesising Subtasks Using the Hierarchical Model

The OLID dataset was used to perform cross-dataset training and testing similar to the experiments of the previous section. However, since OLID uses a hierarchical annotation model (differing from the annotation schemes of the other datasets), this task was approached from a different angle. A model trained on the three subtasks of the OLID dataset (described at the end of Section 3.4) was tested for the task of tagging the in-domain Twitter datasets. This makes it possible to not only see how well OLID-trained models generalise to other data, but to identify the overlap between the different subtasks that the other datasets tackle by observing what percentage of documents under each subtask share common OLID tags.

For the OLID classifiers, a BERT model was used without any extra statistical features and with minimal preprocessing (only lower-casing of tweets). The classifiers were then fine-tuned to the different subtasks, again showing a positive correlation between sequence length and classifier performance. For the results to be comparable to those obtained in the OffensEval 2019 shared task, the same test set was used as in that task. Model performances are reported in Table 5, along with what rank the model would have obtained if it had been submitted to OffensEval 2019, showing that the models are competitive when compared to the top shared task submissions.

The tested model was trained for a total of 3 epochs with a batch size of 16 and learning rate e^{-5}. The maximum sequence length was set to 70 for subtasks A and C, but to 60 for subtask B, where over-fitting was observed on sequence length 70. Also in subtask C the model showed significant signs of over-fitting, with the BERT approach only achieving an F_1 score of 0.52. In this case, a technique was borrowed from the top subtask C submission to OffensEval (Radivchev and Nikolov, 2019), namely to use lower decision boundaries for the `OTH` (0.2) and `GRP` (0.3) tags, instead of the typical decision boundary probability of 0.5. As can be seen in the table, this addition led to huge improvements ($F_1 = 0.62$), compared to the models using the typical decision boundary ($F_1 = 0.52$), although the achieved scores still were not close to the top submission.

Returning to the tagging/synthesis experiments, the entire datasets were used. The results are presented in Table 6. Here we see quite a bit of overlap between the offensive and hate speech tags with the majority tag being `(OFF, TIN, IND)` by a landslide. Clearly, these results can become trivial if the differences boil down to whether the model generalises well to the other datasets used here. This is why only in-domain (Twitter datasets) are considered here and the results also are discussed while taking this into account.

In the Davidson et al. dataset, the non-abusive tag, `neither` had a much lower percentage of its tweets annotated under `NOT` (69.37%) by the OLID classifier when compared to other datasets. This observation may be attributed to the data collection techniques used by Davidson et al., who filtered tweets based on a hate speech lexicon before annotating them, as well as to profanities occurring within the `neither` tag, causing a dip in the amount of explicitly non-offensive tweets.

A similar issue is seen, but to a lesser extent, in the `neither` tag of the Waseem and Hovy dataset, which also was extended by using a sample of hateful tweets. Another interesting observation with that dataset is that the majority class for the `sexism` tag in Subtask A was `NOT`. This complies with observations by both Waseem and Hovy and Davidson et al. (2017) that the human coders considered sexist terms as offensive rather than hateful. However, in terms of our classifier, this may only be due to the implicit nature of most sexist insults and a lack of sexist samples within the OLID dataset. Founta et al.'s dataset shows a high number of `hateful` tweets classified as `NOT`, which may be due to the implicit nature of sexism or sarcasm in the tweets involved.

Dataset	Tag	Subtask A		Subtask B		Subtask C		
		OFF	NOT	UNT	TIN	IND	GRP	OTH
Waseem and Hovy	racism	**52.94**	47.06	0.00	**100**	**64.70**	23.52	11.76
	sexism	42.96	**57.04**	8.28	**91.72**	**54.06**	30.27	15.67
	neither	20.22	**79.78**	28.27	**71.73**	**67.11**	25.95	6.94
Davidson et al.	hate_speech	**84.13**	15.87	3.35	**96.65**	**63.22**	24.61	12.17
	offensive_language	**86.89**	13.11	7.45	**92.55**	**71.89**	20.39	7.72
	neither	30.63	**69.37**	20.44	**79.56**	**63.29**	5.48	31.23
Founta et al.	hateful	**78.11**	21.89	5.92	**94.08**	**52.35**	27.37	20.28
	abusive	**97.34**	2.66	26.04	**73.96**	**75.56**	10.42	14.02
	normal	9.23	**90.77**	24.82	**75.18**	**60.24**	35.84	3.92
	spam	8.72	**91.28**	43.59	**56.41**	**50.98**	1.71	47.31
Zampieri et al.	*(actual annotation fraction)*	*32.91*	*67.09*	*11.88*	*88.12*	*61.31*	*28.17*	*10.52*

Table 6: Results of using a BERT model trained on OLID to tag other the datasets, for each OffensEval subtask. The values are percentages of tweets in each class (rows) annotated with the corresponding OLID tag (columns). Note that in the version of the Waseem and Hovy dataset used here, the racism tag only had 17 samples.

Some blanket statements that can be made given these results are that hate speech is highly targeted, mainly at individuals, but with a significant share targeted at groups and other institutions/issues. Offensive language, on the other hand, tends to be highly targeted only at individuals. Furthermore, the dearth of data belonging to the UNT, GRP and OTH tags may have had a detrimental effect on the model leading to the lob-sided (OFF, TIN, IND) classification.

7 Discussion and Conclusion

The paper makes two major contributions: First, an evaluation of the general effectiveness of BERT in Abusive Language Classification tasks and its ability to obtain results comparable to — or better than — the state-of-the-art by only fine-tuning.

Second, experiments showing that datasets with larger percentages of positive samples generalise better than datasets with fewer positive samples when tested against a dissimilar dataset (at least within the same platform, e.g., Twitter), which indicates that a more balanced dataset is healthier for generalisation. This observation should be accounted for when attempting to build new datasets to tackle Abusive Language Detection, but this is far from the only problem faced when attempting to create such datasets.

Looking at the various available datasets in this field, it is obvious that it cannot be expected for a single dataset to encompass all facets of abuse online. For example, on scanning the OLID (Zampieri et al., 2019a) using a lexicon of sexist and racist terms from Hatebase.org only a measly 55 and 567 hits, respectively, were obtained. Armed with this information we cannot possibly expect a model trained on the OLID dataset to effectively detect racism and sexism online. In fact, most of the data in OLID seem to be political, indicating that it in contrast has a high potential to detect such phenomena.

The point made here is that datasets used in the Abusive Language Detection space must be more representative of all facets of abusive language, if we expect them to generalise to any subset of abuse. Also, there are very few datasets that provide a large number of samples that can be taken advantage of by huge neural networks (Lee et al., 2018). However, we do acknowledge the difficulty in collecting abusive samples as most discourse online is benign. To address these issues, all datasets must advertise the subset of the abusive language they represent. In addition, more work must be done to identify similarities and holes in the representation of datasets. Merging of datasets may also prove to be a promising solution to the non-generalisability problem. Waseem et al. (2018)'s multi-task learning model can be a solid starting point for such endeavours.

A more ambitious solution could be the development of pre-trained embeddings (at the word and/or character level) for Abusive Language Detection, although the procurement of enough broad spanning data to produce a high-quality embedding could again be quite a challenging task.

In terms of whether the hierarchical annotation model helps in reducing redundancy and overlap in Abusive Language Detection subtasks, the answer is both yes and no:

- yes, the hierarchical annotation model does reveal the overlap in the subtasks of abusive language detection; but,

- no, it could hardly be a replacement for the existing multi-class annotation schema.

This is because there is still value in identifying whether a sample is racist / sexist / cyberbullying over just recognising whether the abuse is explicit or not, and in identifying the target of abuse.

However, the hierarchical model in its current form still cannot differentiate between various subsets of abusive language. Future hierarchical models could address this either by adding more levels to further differentiate the subsets or by creating additional levels to identify subsets more explicitly. For example, after the first level of the OLID (Zampieri et al., 2019a) annotation schema, it could branch out into a layer that classifies samples as hate speech, bullying / trolling or as non-abusive use of offensive language. The hate speech tag could then be expanded into another level classifying hate speech as being, e.g., racism, sexism, or other. This way of moving from coarse-grained tags to increasingly finer-grained ones might be a workable approach to tackling hierarchical annotation.

Other issues such as the adversarial methods used to bypass detection methods (Gröndahl et al., 2018) also plague this problem space. Character-based features alleviate this complication to some degree, but more work needs to be done to solve this. Research in this domain has also largely constrained itself to text, while real-world scenarios are quite different — there is a huge section of abuse online that rely on other forms of communication such as images, videos and gifs.

An overall conclusion is that the data is more important than the model when tackling Abusive Language Detection. Schmidt and Wiegland (2017) expressed the need for a benchmark dataset for abusive language tasks, but it would be unwise to say any current dataset fills this role. Future work must focus more on how models generalise to the real world by modifying the testing procedure. A model's performance on the dataset it was trained on cannot be indicative of how well it would perform in a real-life application, and a dataset's quality must be measured on how broad spanning and how representative it is of abusive language as a whole.

Acknowledgments

Thanks to all the researchers who have made their datasets available, specially Waseem and Hovy, Davidson et al., Founta et al., and Zampieri et al., the organisers of SemEval-2019 Task 6: OffensEval ('Identifying and Categorizing Offensive Language in Social Media').

Special thanks to the anonymous reviewers whose comments helped to improve the paper.

References

Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval: 35th European Conference on IR Research*, pages 693–696. Springer, Moscow, Russia.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. ACL.

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 Conference on Web and Social Media*, pages 512–516, Montréal, Canada. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web: Papers from the 2011 ICWSM Workshop*, pages 11–17, Barcelona, Spain. AAAI Press.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30, Florence, Italy. ACM.

Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont, editors. 2018. *Proceedings of the 2nd Workshop on Abusive Language Online*. ACL, Brussels, Belgium.

Paula Fortuna, José Ferreira, Luiz Pires, Guilherme Routar, and Sérgio Nunes. 2018. Merging datasets for aggressive text identification. In (Kumar et al., 2018b), pages 128–139.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. *CoRR*, abs/1802.00393.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In (Waseem et al., 2017a), pages 85–90.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the 11th International Conference on Recent Advances in Natural Language Processing*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM Web on Science Conference*, pages 229–233, Troy, New York, USA. ACM.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 2–12, Toronto, Canada. ACM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the Instagram social network. *CoRR*, abs/1503.03909.

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In (Fišer et al., 2018), pages 132–137.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2018. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. Manuscript, Simon Fraser University, Vancouver, Canada.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In (Kumar et al., 2018b), pages 1–11.

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018b. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*. ACL, Santa Fe, New Mexico, USA.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on Twitter. In (Fišer et al., 2018), pages 101–106.

Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors. 2019. *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval)*. ACL, Minneapolis, Minnesota, USA.

Yashar Mehdad and Joel R. Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, Los Angeles, California, USA. ACL/SIGDIAL.

Johannes Skjeggestad Meyer and Björn Gambäck. 2019. A platform agnostic dual-strand hate speech detector. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 146–156, Florence, Italy. ACL.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, Montréal, Canada. IW3C2.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Georgios Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):47304742.

Victor Radivchev and Alex Nikolov. 2019. Nikolov-Radivchev at SemEval-2019 Task 6: Offensive tweet classification with BERT and ensembles. In (May et al., 2019), pages 687–691.

Sarah T. Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven, Connecticut, USA.

Anna Schmidt and Michael Wiegland. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. ACL.

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. *CoRR*, abs/1603.07709.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on Twitter hate speech detection. In (Fišer et al., 2018), pages 75–86.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the 10th Conference on Recent Advances in Natural Language Processing, Proceedings*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the World Wide Web. In *Proceedings of the 2nd Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. ACL.

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors. 2017a. *Proceedings of the First Workshop on Abusive Language Online*. ACL, Vancouver, Canada.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017b. Understanding abuse: A typology of abusive language detection subtasks. In (Waseem et al., 2017a), pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the North American Chapter of the Association for Computational Linguistics, Student Research Workshop*, pages 88–93, San Diego, California, USA. ACL.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In Jennifer Golbeck, editor, *Online Harassment*, pages 29–55. Springer, Cham, Switzerland.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, Perth, Australia. IW3C2.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada. ACL.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. 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), Minneapolis, Minnesota, USA. ACL.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In (May et al., 2019), pages 75–86.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Proceedings of the 15th International Semantic Web Conference*, pages 745–760, Heraklion, Greece. Springer.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the Instagram social network. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3952–3958, New York, New York, USA. AAAI Press.