A Large-Scale English Multi-Label Twitter Dataset for Online Abuse Detection

Semiu Salawu
Aston University
Birmingham B4 7ET
salawusd@aston.ac.uk

Prof. Jo Lumsden
Aston University
Birmingham B4 7ET
j.lumsden@aston.ac.uk

Prof. Yulan He
The University of Warwick
Coventry CV4 7AL
Yulan.He@warwick.ac.uk

Abstract

In this paper, we introduce a new English Twitter-based dataset for online abuse and cyberbullying detection. Comprising 62,587 tweets, this dataset was sourced from Twitter using specific query terms designed to retrieve tweets with high probabilities of various forms of bullying and offensive content, including insult, profanity, sarcasm, threat, porn and exclusion. Analysis performed on the dataset confirmed common cyberbullying themes reported by other studies and revealed interesting relationships between the classes. The dataset was used to train a number of transformer-based deep learning models returning impressive results.

1 Introduction

Cyberbullying has been defined as an aggressive and intentional act repeatedly carried out using electronic means against a victim that cannot easily defend him or herself (Smith et al., 2008). Online abuse by contrast can refer to a wide range of behaviours that may be considered offensive by the parties to which it is directed to (Sambaraju and McVittie, 2020). This includes trolling, cyberbullying, sexual exploitation such as grooming and sexting and revenge porn (i.e., the sharing of inappropriate images of former romantic partners). A distinguishing feature of cyberbullying within the wider realm of online abuse is that it is a repeated act and its prevalence on social media (along with other forms of online abuse) has generated significant interest in its automated detection. This has lead to an increase in research efforts utilising supervised machine learning methods to achieve this automated detection. Training data plays a significant role in the detection of cyberbullying and online abuse. The domain-bias, composition and taxonomy of a dataset can impact the suitability of models trained on it for abuse detection purposes, and therefore the choice of training data plays a significant role in the performance of these tasks.

While profanity and online aggression are often associated with online abuse, the subjective nature of cyberbullying means that accurate detection extends beyond the mere identification of swear words. Indeed, some of the most potent abuse witnessed online has been committed without profane or aggressive language. This complexity often requires labelling schemes that are more advanced than the binary annotation schemes used on many existing labelled datasets. This, therefore, influenced our approach in creating the dataset. After extracting data from Twitter using targeted queries, we created a taxonomy for various forms of online abuse and bullying (including subtle and indirect forms of bullying) and identified instances of these and other inappropriate content (e.g., pornography and spam) present within the tweets using a fine-grained annotation scheme. The result is a large labelled dataset with a majority composition of offensive content.

This paper is organised as follows. In Section 2, we present an overview of existing online abuse-related datasets. Section 3 discusses the collection method, composition, annotation process and usage implications for our dataset. Results of the experiments performed using the dataset are discussed in Section 4. Finally, conclusion and future research are described in section 5.

2 Related Work

Social media has become the new playground and, much like physical recreation areas, bullies inhabit facets of this virtual world. The continually evolving nature of social media introduces a need for datasets to evolve in tandem to maintain relevance. Datasets such as the Barcelona Media dataset used in studies such as those by Dadvar and Jong (2012),

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 146–156

Nahar et al. (2014), Huang et al. (2014), Nandhini and Sheeba (2015) was created over ten years ago and, while representative of social media usage at the time, social networks such as Myspace, Slashdot, Kongregate and Formspring from which some of the data was sourced are no longer widely used. The consequence of this is that such datasets are no longer representative of current social media usage. Twitter is one of the most widely used social media platforms globally; as such, it is no surprise that it is frequently used to source cyberbullying and online abuse data.

Bretschneider et al. (2014) annotated 5,362 tweets, 220 of which were found to contain online harassment; the low proportion of offensive tweets present within the dataset (less than 0.05%), however, limits its efficacy for classifier training. More recently, studies such as those by Rajadesingan et al. (2015), Waseem and Hovy (2016), Davidson et al. (2017), Chatzakou et al. (2017), Hee et al. (2018), Founta et al. (2018), Ousidhoum et al. (2019) have produced datasets with higher positive samples of cyberbullying and online abuse.

Rajadesingan et al. (2015) labelled 91,040 tweets for sarcasm. This is noteworthy because while sarcasm can be used to perpetrate online bullying, it is rarely featured in existing cyberbullying datasets' taxonomies. However, as the dataset was created for sarcasm detection only, this is the only context that can be learned from the dataset. As such, any model trained with this dataset will be unable to identify other forms of online abuse, thus limiting its usefulness. Waseem and Hovy (2016) annotated 17,000 tweets using labels like racism and sexism, and Davidson et al. (2017) labelled over 25,000 tweets based on the presence of offensive and hate speech. Chatzakou et al. (2017) extracted features to identify cyberbullies by clustering 9,484 tweets attributed to 303 unique Twitter users. In creating their bi-lingual dataset sourced from ASKfm, Hee et al. (2018) used a detailed labelling scheme that acknowledges the different types of cyberbullying discovered in the retrieved post types. The dataset's effectiveness in training classifiers may, however, be affected by the low percentage of abusive documents present. This dataset was subsequently re-annotated by Rathnayake et al. (2020) to identify which of the four roles of 'harasser', 'victim', 'bystander defender' and 'bystander assistant' was played by the authors of the posts contained in the dataset. Similarly, Sprugnoli et al. (2018) used the same four roles to annotate a dataset created from simulated cyberbullying episodes using the instant messaging tool; WhatsApp, along with the labels created by Hee et al. (2018)

Zampieri et al. (2019) used a hierarchical annotation scheme that, in addition to identifying offensive tweets, also identifies if such tweets are targeted at specific individuals or groups and what type of target it is (i.e., individual - *@username* or group – '*... all you republicans*'). Hierarchical annotation schemes have indeed shown promise as observed in their use in recent offensive language detection competitions like hatEval[1] and OffensEval[2]; that said, however, a hierarchical scheme could inadvertently filter out relevant labels depending on the first-level annotation scheme used.

Ousidhoum et al. (2019) used one of the most comprehensive annotation schemes encountered in an existing dataset and additionally included a very high percentage of positive cyberbullying samples in their dataset but, regrettably, the number of English documents included in the dataset is small in comparison to other datasets. Founta et al. (2018) annotated about 10,000 tweets using labels like abusive, hateful, spam and normal, while Bruwaene et al. (2020) experimented with a multi-platform dataset comprising of 14,900 English documents sourced from Instagram, Twitter, Facebook, Pinterest, Tumblr, YouTube and Gmail. Other notable publicly available datasets include the Kaggle Insult (Kaggle, 2012) and Kaggle Toxic Comments (Kaggle, 2018) datasets. A comprehensive review of publicly available datasets created to facilitate the detection of online abuse in different languages is presented in Vidgen and Derczynski (2020).

3 Data

In this section, we introduce our dataset and how it addresses some of the limitations of existing datasets used in cyberbullying and online abuse detection research.

3.1 Objective

In reviewing samples of offensive tweets from Twitter and existing datasets, we discovered that a single tweet could simultaneously contain elements of abuse, bullying, hate speech, spam and many other forms of content associated with cyberbullying. As such, attributing a single label to a tweet ignores

[1] competitions.codalab.org/competitions/19935
[2] sites.google.com/site/offensevalsharedtask

Label	Description	Example
Bullying	Tweets directed at a person(s) intended to provoke and cause offence. The target of the abuse must be from the tweet either via mentions or names.	@username You are actually disgusting in these slutty pictures Your parents are probably embarrassed...
Insult	Tweets containing insults typically directed at or referencing specific individual(s).	@username It's because you're a c*nt isn't it? Go on you are aren't you?
Profanity	This label is assigned to any tweets containing profane words.	@username please dont become that lowkey hating ass f**king friend please dont
Sarcasm	Sarcastic tweets aimed to ridicule. These tweets may be in the form of statements, observations and declarations.	@username Trump is the most innocent man wrongly accused since O.J. Simpson. #Sarcasm
Threat	Tweets threatening violence and aggression towards individuals.	@username Let me at him. I will f*ck him up and let my cat scratch the f*ck out of him.
Exclusion	Tweets designed to cause emotional distress via social exclusion.	@username @username You must be gay huh ? Why you here ? Fag !! And I got 2 TANK YA !
Porn	Tweets that contain or advertise pornographic content	CLICK TO WATCH [link] Tinder Sl*t Heather Gets her A*s Spanks and Spreads her C*nt
Spam	Unsolicited tweets containing and advertising irrelevant content. They typically include links to other web pages	HAPPY #NationalMasturbationDay #c*m and watch me celebrate Subscribe TODAY for a free #p*ssy play video of me [link]

Table 1: Annotation scheme with examples.

other salient labels that can be ascribed to the tweet. We propose a multi-label annotation scheme that identifies the many elements of abusive and offensive content that may be present in a single tweet. As existing cyberbullying datasets often contain a small percentage of bullying samples, we want our dataset to contain a sizeable portion of bullying and offensive content and so devised querying strategies to achieve this. Twitter, being one of the largest online social networks with a user base in excess of 260 million (Statista, 2019) and highly representative of current social media usage, was used to source the data.

3.2 Labels

Berger (2007) (as cited in Abeele and Cock 2013, p.95) distinguishes two types of cyberbullying, namely direct and indirect/relational cyberbullying. Direct cyberbullying is when the bully directly targets the victim (typified by sending explicit offensive and aggressive content to and about the victim) while indirect cyberbullying involves subtler forms of abuse such as social exclusion and the use of sarcasm to ridicule. As both forms of cyberbullying exist on Twitter, our annotation scheme (see Table 1) was designed to capture the presence of both forms of bullying within tweets.

3.3 Collection Methods

Offensive and cyberbullying samples are often minority classes within a cyberbullying dataset; as such, one of our key objectives was ensuring the inclusion of a significant portion of offensive and cyberbullying samples within the dataset to facilitate training without the need for oversampling. Rather than indiscriminately mining Twitter feeds, we executed a series of searches formulated to return tweets with a high probability of containing the various types of offensive content of interest. For insulting and profane tweets, we queried Twitter using the 15 most frequently used profane terms on Twitter as identified by Wang et al. (2014). These are: f*ck, sh*t, a*s, bi*ch, ni**a, hell, wh*re, d*ck, p*ss, pu**y, sl*t, p*ta, t*t, damn, f*g, c*nt, c*m, c*ck, bl*wj*b, retard. To retrieve tweets containing sarcasm, we used a strategy based on the work of Rajadesingan et al. (2015) which discovered that sarcastic tweets often include *#sarcasm* and *#not* hashtags to make it evident that sarcasm was the intention. For our purposes, we found *#sarcasm*

more relevant and therefore queried Twitter using this hashtag.

To discover prospective query terms for threatening tweets, we reviewed a random sample of 5000 tweets retrieved via Twitter's Streaming API and identified the following hashtags as potential query terms: *#die*; *#killyou*; *#rape*; *#chink*, *#muslim*, *#FightAfterTheFight* and *#cops*. These hashtags were then used as the initial seed in a snowballing technique to discover other relevant hashtags. This was done by querying Twitter using the hashtags and inspecting the returned tweets for violence-related hashtags. The following additional hashtags were subsequently discovered through this process: *#killallblacks*; *#killallcrackers*; *#blm*; *#blacklivesmatter*; *#alllivesmatter*; *#bluelivesmatter*; *#killchinese*; *#bustyourhead*; *#f*ckyouup*; *#killallwhites*; *#maga*; *#killallniggas*; and *#nigger*.

Formulating a search to retrieve tweets relating to social exclusion was challenging as typical examples were rare. From the 5000 tweets seed sample, we classified six tweets as relating to social exclusion and from them identified the following hashtags for use as query terms: *#alone*, *#idontlikeyou* and *#stayinyourlane*. Due to the low number of tweets returned for these hashtags, we also extracted the replies associated with the returned tweets and discovered the following additional hashtags *#notinvited*, *#dontcometomyparty*, and *#thereisareasonwhy* which were all subsequently used as additional query terms. Rather than excluding re-tweets when querying as is common practice amongst researchers, our process initially extracted original tweets and retweets and then selected only one of a tweet and its retweets if they were all present in the results. This ensured relevant content was not discarded in situations where the original tweet was not included in the results returned, but retweets were. Our final dataset contained 62,587 tweets published in late 2019.

3.4 Annotation Process

Language use on social media platforms like Twitter is often colloquial; this, therefore, influenced the desired annotator profile as that of an active social media user that understands the nuances of Twitter's colloquial language use. While there is no universal definition of what constitutes an active user on an online social network, Facebook defined an active user as someone who has logged into the site and completed an action such as liking, sharing and posting within the previous 30 days (Cohen, 2015). With one in every five minutes spent online involving social media usage and an average of 39 minutes spent daily on social media (Ofcom Research, 2019), this definition is inadequate in view of the increased users' activities on social media. An active user was therefore redefined as one that has accessed any of the major social networks (e.g., Twitter, Instagram, Facebook, Snapchat) at least twice a week and made a post/comment, like/dislike or tweet/retweet at least once in the preceding two weeks. This new definition is more in keeping with typical social media usage.

Using personal contacts, we recruited a pool of 17 annotators. Our annotators are from different ethnic/racial backgrounds (i.e., Caucasian, African, Asian, Arabian) and reside in different countries (i.e., US, UK, Canada, Australia, Saudi Arabia, India, Pakistan, Nigeria and Ghana). Additionally, their self-reported online social networking habits met our definition of an active social media user. All annotators were provided with preliminary information about cyberbullying including news articles and video reports, documentaries and YouTube videos as well as detailed information about the labelling task. Due to the offensive nature of the tweets and the need to protect young people from such content while maintaining an annotator profile close to the typical age of the senders and recipients of the tweets, our annotators were aged 18 - 35 years.

Since the presence of many profane words can be automatically detected, a program was written to label the tweets for profane terms based on the 15 profane words used as query terms and the Google swear words list[3]. The profanity-labelled tweets were then provided to the annotators to alleviate this aspect of the labelling task. Each tweet was labelled by three different annotators from different ethnic/racial backgrounds, gender and countries of residence. This was done to control for annotators' cultural and gender bias.

An interesting observation of the annotation process was the influence of the annotators' culture on how labels are assigned. For example, we discovered that annotators from Asian, African and Arabian countries were less likely to assign the 'bullying', 'insult' and 'sarcasm' labels to tweets compared to annotators from the UK, Canada, US

[3] code.google.com/archive/p/badwordslist/downloads

and Australia. A possible explanation for this could be that the context of the abuse apparent to the annotators from the Caucasian countries may not translate well to other cultures. While no other substantial trend were noticed for the other labels, this, however, highlighted the impact of an annotator's personal views and culture on the labelling task and the labels' composition of our dataset could have been different if we had sourced annotators from different cultures. As identified by Bender and Friedman (2018), researchers should therefore be mindful of potential annotators' biases when creating online abuse datasets.

Inter-rater agreement was measured via Krippendorff's Alpha (α) and the majority of annotators' agreement was required for each label. The Krippendorff python library[4] was used to compute the value which was found to be 0.67 which can be interpreted as 'moderate agreement'. We believe that the culturally heterogeneous nature of our annotators pool could have 'diluted' the agreement amongst annotators and contributed to the final value achieved.

3.5 Analysis

The number of tweets each label was assigned to is presented in Table 2 with 'Profanity' emerging as the dominant label and 'Exclusion' the least assigned label. It can also be seen that about a sixth of the tweets were not assigned any labels.

Label	Profanity	Porn	Insult
Count	51,014	16,690	15,201
Label	Spam	Bullying	Sarcasm
Count	14,827	3,254	117
Label	Threat	Exclusion	None
Count	79	10	10,768

Table 2: Number of tweets each label was assigned to.

Before preprocessing, the maximum document length for the dataset was 167 characters with an average document length of 91. Following preprocessing, the maximum document length reduced to 143 characters (equating to 26 words) with an average document length of 67 characters. The removal of mentions (i.e., including a username with the @ symbol inside a tweet), URLs and non-ASCII characters were found to be the biggest contributor to document length reduction. There are 37,453 unique tokens in the dataset. Figure 1 illustrates the number of tweets assigned to multiple labels. Single label tweets make up more than a third of the dataset, which can be mostly attributed to the large number of tweets singly labelled as 'Profanity'.

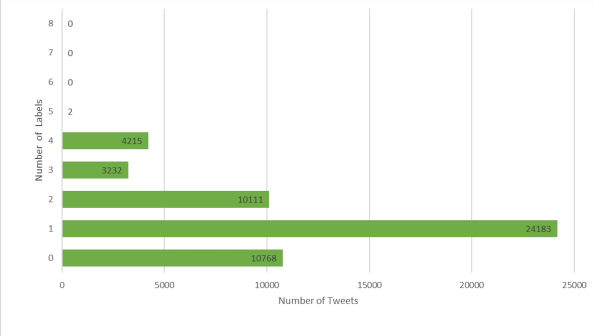

Figure 1: Distribution of tweet counts and number of labels assigned.

A significant number of tweets were also jointly labelled as 'Profanity' and 'Insult' or 'Insult' and 'Cyberbullying', and this contributed to double-labelled tweets being the second-largest proportion of the dataset. Interestingly, there were more tweets with quadruple labels than there were with triple and this was discovered to be due to the high positive correlation between 'Porn'/'Spam' and 'Profanity'/'Insult.'

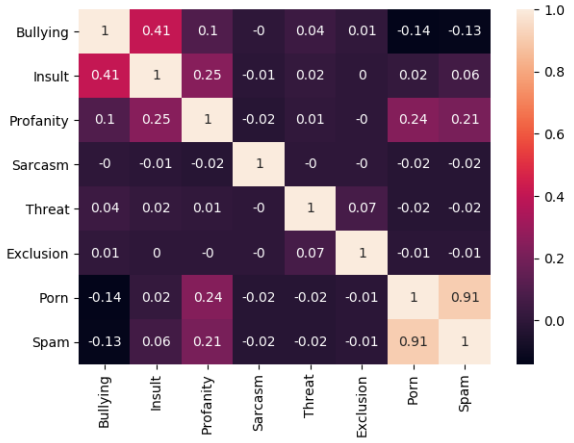

Figure 2: Correlation matrix for dataset's labels.

The correlation matrix for the classes in the dataset is illustrated in Figure 2. The closer the correlation value is to 1, the higher the positive correlation between the two classes. The highest positive correlation is shown to be between 'Porn' and 'Spam' (0.91) followed by 'Insult' and 'Bullying' (0.41) and 'Insult' and 'Profanity' (0.25). 'Porn' and 'Spam' also demonstrated a positive correlation between them and 'Profanity' which can be

[4] pypi.org/project/krippendorff

attributed to the high proportion of profane terms in pornographic content and spam; we found that many pornographic tweets are essentially profanity-laden spam. 'Insult' also exhibited a positive correlation with 'Bullying' and 'Profanity', a fact that can be attributed to the frequent use of profanity in insulting tweets as well as the use of insults to perpetrate bullying. The key negative correlations identified by the chart includes those between 'Bullying', and 'Porn' and 'Spam'. This can be attributed to bullying tweets often being personal attacks directed at specific individuals and typified by the use of usernames, person names or personal pronouns, all of which are rare in pornographic and spam tweets. The minority classes 'Sarcasm', 'Threat' and 'Exclusion' exhibited a minimal correlation with the other classes.

3.6 Bias Implication

Most datasets carry a risk of demographic bias (Hovy and Spruit, 2016) and this risk can be higher for datasets created using manually-defined query terms. Researchers, therefore, need to be aware of potential biases in datasets and address them where possible. Gender and ethnicity are common demographic biases that can be (often inadvertently) introduced into a dataset. To this end, we wanted to explore (as far as possible), whether our dataset had acquired gender bias. To do this we attempted to infer the gender of the users incorporated in our dataset. Since Twitter does not record users' gender information, we adopted an approach that uses the Gender API [5] to deduce the gender of users based on whether the users' first names are traditionally male or female: we assumed that as an accessible and feasible measure of users' gender identity. We were able to process the authorship of 13,641 tweets (21.8% of the dataset) in this way and inferred that 31.4% of the authors of these tweets identified as female and 68.6% male (at least in so far as was apparent from their Twitter account). This suggests a male-bias in the authorship of the tweets in the dataset. We, however, recognise the limitation of this approach as the names provided by users cannot always be regarded as truthful and as gender extends beyond the traditional binary types, a names-based approach such as this cannot be used to deduce all gender identities. A more empathetic and effective means to identify gender in Twitter users would be an interesting facet of future work.

With regards racial and ethnic bias, we mitigate potential bias by including generalised variants of any ethnicity-specific keyword used as a query term as well as including variants for different ethnicities. It should, however, be noted that the popularity and topicality of certain keywords may still introduce an unintended bias. For example, #blacklivematters returns several more tweets than #asianlivematters.

While the collection strategy used to create our dataset ensured a high concentration of offensive tweets, a potential consequence of the imbalanced distribution of the classes is that it may reinforce the unintentional bias of associating minority classes to specific hateful and offensive content. Dixon et al. (2018) defined unintended bias as when a model performs better for comments containing specific terms over others. For example, the phrase 'stay in your lane' was found in 4 of the 10 tweets identified as 'Exclusion' (due to the use of the hashtag #stayinyourlane as a query term), this can cause a model trained on the dataset to overgeneralised the phrase's association with the 'Exclusion' label, thus introducing a false positive bias in the model. Introducing more examples of the minority classes using a variety of query terms is a potential strategy for mitigating such unintended bias and is discussed further under future work.

3.7 Practical Use

Ultimately the aim of a dataset such as this is to train machine learning models that can subsequently be used in abuse detection systems. It is, therefore, crucial to understand how any bias in the dataset is manifested in the trained model and the impact of such bias in practical applications. A National Institute of Science and Technology (NIST) study (Grother et al., 2019) discovered that, for example, many US-developed facial recognition algorithms generated significantly higher false positives for Asian and African-American faces compared to Caucasian faces while similar algorithms developed in Asian countries did not show any such dramatic differences in false positive rates between Asian, African-American and Caucasian faces. The study concluded that the use of diverse training data is critical to the reduction of bias in such AI-based applications.

Our dataset has been used to train the classifier used in an online abuse prevention app (called

[5] https://gender-api.com

BullStop) which is available to the public via the Google play store. The app detects offensive messages sent to the user and automatically deletes them. It, however, acknowledges the possibility of both false positive and negative predictions, and thus allows the user to review and re-classify deleted messages and uses such corrections to re-train the system. This is especially important for a subjective field such as online abuse detection.

4 Experiments

4.1 Setup

Models for comparison We experimented with both traditional classifiers (Multinomial Naive Bayes, Linear SVC, Logistic Regression) and deep learning-based models (BERT, Roberta, XLNet, DistilBERT) to perform multi-label classification on the dataset. BERT (Bidirectional Encoder Representations from Transformers) is a language representation model designed to pre-train deep bi-directional representations from unlabeled text (Devlin et al., 2019). RoBERTa (Robustly Optimized BERT Pretraining Approach) is an optimised BERT-based model (Liu et al., 2019), and DistilBERT (Distilled BERT) is a compacted BERT-based model (Sanh et al., 2019) that requires fewer computing resources and training time than BERT (due to using about 40% fewer parameters) while preserving most of BERT performance gains. XLNet (Yang et al., 2019) is an autoregressive language model designed to overcome some of the limitations of BERT. BERT, RoBERTa, XLNet, and DistilBERT are available as pre-trained models but can also be fine-tuned by first performing language modelling on a dataset.

Evaluation Each model's performance was evaluated using macro ROC-AUC (Area Under ROC Curve), Accuracy, Hamming Loss, Macro and Micro F_1 Score, which are typically used in imbalanced classification tasks.

4.2 Preprocessing

The primary objective of our preprocessing phase was the reduction of irrelevant and noisy data that may hamper classifier training. As is standard for many NLP tasks, punctuation, symbols and non-ASCII characters were removed. This was followed by the removal of mentions and URLs. We also discovered many made-up words created by combining multiple words (e.g. goaway, itdoesntwork, gokillyourself) in the tweets. These are due to hashtags, typos and attempts by users to mitigate the characters limit imposed by Twitter. The wordsegment python library was used to separate these into individual words. The library contains an extensive list of English words and is based on Google's 1T (1 Trillion) Web corpus.[6] Lastly, the text was converted to lower case.

4.3 Results

We provide the stratified 10-fold cross-validation results of the experiments in Table 3. The best macro ROC-AUC score was achieved by the pre-trained RoBERTa model, while the best macro and micro F_1 scores were attained using the pre-trained BERT and RoBERTa models, respectively. The best overall accuracy was returned by the fine-tuned DistilBERT model. As expected, the deep learning models outperformed the baseline classifiers with Multinomial Naive Bayes providing the worst results across the experiments and the BERT-like models achieving the best results for each metric. Interestingly, the pre-trained models were marginally better than the equivalent fine-tuned models implying that fine-tuning the models on the dataset degrades rather than improves performance.

As would be expected, the models performed better at predicting labels with higher distributions. For the minority classes like Sarcasm, Threat and Exclusion, RoBERTA and XLNet performed better. All the models performed well in predicting the none class, i.e. tweets with no applicable labels.

The resulting dataset from our collection methods is imbalanced with a high percentage of cyberbullying tweets. In reality, such a concentration of cyberbullying and offensive tweets is highly unusual and at odds with other cyberbullying datasets. To evaluate the generalisability of models trained on our dataset, we performed further experiments to evaluate how the models perform on other unseen datasets. We used our best performing model; RoBERTa (pre-trained), to perform prediction on samples extracted from two other datasets and compared the results against that achieved on our dataset by RoBERTa models trained on the other datasets.

The dataset created by Davidson et al. (2017) and the Kaggle Toxic Comments dataset (Kaggle, 2018) were selected for the experiments. We re-

[6] https://catalog.ldc.upenn.edu/LDC2006T13.

Model	Macro ROC-AUC(↑)	Accuracy (↑)	Hamming Loss (↓)	Macro F_1(↑)	Micro F_1(↑)
Multinomial Naive Bayes	0.8030	0.4568	0.1014	0.2618	0.7200
Linear SVC	0.8353	0.5702	0.0866	0.3811	0.7674
Logistic Regression	0.8354	0.5743	0.0836	0.3587	0.7725
BERT (pre-trained)	0.9657	0.5817	0.0736	**0.6318**	0.7998
DistilBERT (pre-trained)	0.9675	0.5802	0.0764	0.5202	0.7855
RoBERTa (pre-trained)	**0.9695**	0.5785	**0.0722**	0.5437	**0.8081**
XLNet(pre-trained)	0.9679	0.5806	0.0738	0.5441	0.8029
BERT (fine-tuned)	0.9651	0.5822	0.0725	0.5300	0.8022
DistilBERT (fine-tuned)	0.9633	**0.5834**	0.0753	0.5040	0.7872
RoBERTa (fine-tuned)	0.9670	0.5794	0.0724	0.5329	0.8044
XLNet(fine-tuned)	0.9654	0.5819	0.0741	0.5308	0.8037

Table 3: Results of classification. (↑: higher the better; ↓: lower the better)

Model	Macro ROC-AUC(↑)	Accuracy (↑)	Hamming Loss (↓)	Macro F_1(↑)	Micro F_1(↑)
$\text{RoBERTa}_{C\rightarrow D}$	**0.9923**	0.8809	**0.0288**	**0.8802**	**0.8810**
$\text{RoBERTa}_{D\rightarrow C}$	0.9681	0.5831	0.0708	0.5330	0.8076
$\text{RoBERTa}_{D\rightarrow D}$	0.9905	**0.8814**	0.0300	0.8427	0.8758
$\text{RoBERTa}_{C\rightarrow K}$	**0.9916**	0.5924	**0.0123**	**0.5670**	0.7436
$\text{RoBERTa}_{K\rightarrow C}$	0.9651	0.5811	0.0727	0.5352	**0.8054**
$\text{RoBERTa}_{K\rightarrow K}$	0.9733	**0.8449**	0.0174	0.5026	0.6354

Table 4: Results of cross-domain experiments. (↑: higher the better; ↓: lower the better)

ferred to these as the Davidson (D) and the Kaggle (K) datasets and our dataset as the Cyberbullying (C) dataset. The Davidson dataset is a multi-class-labelled dataset sourced from Twitter where each tweet is labelled as one of 'hate_speech', 'offensive' and 'neither'. In contrast, the Kaggle datasets contained Wikipedia documents labelled using a multi-label annotation scheme with each document associated with any number of classes from 'toxic','severe_toxic', 'obscene', 'threat', 'insult' and 'identity_hate'. Due to the difference in the number of labels for each dataset (our dataset contained 8 labels while the Davidson and Kaggle datasets used 3 and 6 labels respectively), it was necessary to amend the input and output layers of the RoBERTa model to allow it to predict the relevant labels for the Davidson and Kaggle datasets

We evaluated our model on the Davidson and Kaggle datasets and for the reverse experiments, evaluated new instances of RoBERTa trained on the other datasets on samples of the Cyberbullying dataset. As control experiments, RoBERTa models were trained and evaluated on the other datasets. The results of our experiments are presented in Table 4.

Overall, models trained on our dataset ($\text{RoBERTa}_{C\rightarrow D}$ and $\text{RoBERTa}_{C\rightarrow K}$) perform better on the other two datasets than the models trained on the other datasets and tested on the Cyberbullying dataset ($\text{RoBERTa}_{D\rightarrow C}$, $\text{RoBERTa}_{K\rightarrow C}$). Interestingly, models trained on our dataset achieved better ROC-AUC, Macro and Micro F_1 values on both the Davidson (D) and the Kaggle (K) datasets compared to in-domain results on those datasets (i.e., models trained and evaluated on the same datasets - $\text{RoBERTa}_{D\rightarrow D}$ and $\text{RoBERTa}_{K\rightarrow K}$). The results indicate that our dataset sufficiently captures enough context for classifiers to distinguish between both cyberbullying and non-cyberbullying text across different

social media platforms.

4.4 Discussion and Future Work

Our collection strategy for creating the dataset was designed to target cyberbullying and offensive tweets and ensure that these types of tweets constitute the majority class. This differs from the collection strategies used in other datasets such as those by Dadvar et al. (2013), Kontostathis et al. (2013) and Hosseinmardi et al. (2015) which are designed to simulate a more realistic distribution of cyberbullying. As the occurrence of cyberbullying documents is naturally low, classifiers trained on our dataset can benefit from a high concentration of cyberbullying and offensive documents without the need for oversampling techniques.

When cross-domain evaluation was performed using our best performing classifier on two other datasets (Davidson et al., 2017; Kaggle, 2018), the model trained on our dataset performed better than those trained on the other datasets. It is also worth noting that the composition and annotation of these other datasets is entirely different from ours, and one was sourced from a different platform (Wikipedia). Our results demonstrated that deep learning models could learn sufficiently from an imbalanced dataset and generalise well on different data types.

We discovered a slight performance degradation for the deep learning-based models after fine-tuning. As recently shown in (Radiya-Dixit and Wang, 2020), fine-tuned networks do not deviate substantially from pre-trained ones and large pre-trained language models have high generalisation performance. We will explore in future work, more effective ways for producing fine-tuned networks such as learning to sparsify pre-trained parameters and optimising the most sensitive task-specific layers.

The distribution of 'Sarcasm', 'Exclusion' and 'Threat' labels is low within the dataset. Consequently, the models' ability to predict these classes is not comparable to that of the majority classes. Increasing the distribution of these labels within the dataset will improve the models training and mitigate unintended bias that may have been introduced by the minority classes; we therefore plan to supplement the dataset with more positive samples of these classes by exploring other querying strategies as well as incorporating samples from existing datasets such as Rajadesingan et al. (2015) and Hee et al. (2018).

5 Conclusion

In this paper, we presented a new cyberbullying dataset and demonstrated the use of transformer-based deep learning models to perform fine-grained detection of online abuse and cyberbullying with very encouraging results. To our knowledge, this is the first attempt to create a cyberbullying dataset with such a high concentration (82%) of cyberbullying and offensive content in this manner and using it to successfully evaluate a model trained with the dataset on a different domain. The dataset is available at https://bitbucket.org/ssalawu/cyberbullying-twitter for the use of other researchers.

References

Mariek Vanden Abeele and Rozane De Cock. 2013. Cyberbullying by mobile phone among adolescents: The role of gender and peer group status. *Communications*, 38:107–118.

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.

Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. 2014. Detecting online harassment in social networks.

David Van Bruwaene, Qianjia Huang, and Diana Inkpen. 2020. A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, pages 1–24.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. pages 13–22. Association for Computing Machinery, Inc.

David Cohen. 2015. Facebook changes definition of monthly active users.

Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: A step toward a safer internet yard. pages 121–125.

Maral Dadvar, Dolf Trieschnigg, and Franciska De Jong. 2013. Expert knowledge for automatic detection of bullies in social networks. pages 57–64.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. volume 7, pages 67–73. Association for Computing Machinery, Inc.

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.

Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. 2019. Face recognition vendor test (frvt) part 3: Demographic effects.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PLOS ONE*, 13:e0203794.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Poster: Detection of cyberbullying in a mobile social network: Systems issues. page 481. Association for Computing Machinery, Inc.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. pages 591–598.

Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. pages 3–6. ACM.

Kaggle. 2012. Detecting insults in social commentary.

Kaggle. 2018. Toxic comment classification challenge.

April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: Query terms and techniques. volume volume, pages 195–204. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang. 2014. Semi-supervised learning for cyberbullying detection in social networks. volume 8506 LNCS, pages 160–171. Springer Verlag.

B. Sri Nandhini and J. I. Sheeba. 2015. Online social network bullying detection using intelligence techniques. volume 45, pages 485–492. Elsevier B.V.

Ofcom Research. 2019. Online nation. *Ofcom Research*.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Evani Radiya-Dixit and Xin Wang. 2020. How fine can fine-tuning be? learning efficient language models.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter:a behavioral modeling approach. pages 97–106. Association for Computing Machinery, Inc.

Gathika Rathnayake, Thushari Atapattu, Mahen Herath, Georgia Zhang, and Katrina Falkner. 2020. Enhancing the identification of cyberbullying through participant roles. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 89–94, Online. Association for Computational Linguistics.

Rahul Sambaraju and Chris McVittie. 2020. Examining abuse in online media. *Social and personality psychology compass*, 14(3):e12521.

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

Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 49:376–385.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

Statista. 2019. Global mobile social penetration rate 2019, by region.

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

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twitter. pages 415–425.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. pages 88–93.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Computing Research Repository*, arXiv:1906.08237.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. volume 1, pages 1415–1420. Association for Computational Linguistics (ACL).