Aggressive language in an online hacking forum

Andrew Caines, Sergio Pastrana, Alice Hutchings & Paula Buttery
Department of Computer Science & Technology
University of Cambridge
Cambridge, U.K.
apc38@cam.ac.uk, sp849@cam.ac.uk,
alice.hutchings@cl.cam.ac.uk, paula.buttery@cl.cam.ac.uk

Abstract

We probe the heterogeneity in levels of abusive language in different sections of the Internet, using an annotated corpus of Wikipedia page edit comments to train a binary classifier for abuse detection. Our test data come from the CrimeBB Corpus of hacking-related forum posts and we find that (a) forum interactions are rarely abusive, (b) the abusive language which does exist tends to be relatively mild compared to that found in the Wikipedia comments domain, and tends to involve aggressive posturing rather than hate speech or threats of violence. We observe that the purpose of conversations in online forums tend to be more constructive and informative than those in Wikipedia page edit comments which are geared more towards adversarial interactions, and that this may explain the lower levels of abuse found in our forum data than in Wikipedia comments. Further work remains to be done to compare these results with other inter-domain classification experiments, and to understand the impact of aggressive language in forum conversations.

1 Introduction

The automatic identification of abusive language online[1] is of growing interest and concerns have proliferated about aggressive Internet behaviours commonly known as 'trolling'. From an applications perspective, the accurate detection of vitriolic language is one of the clearest examples of natural language processing for social good, assuming data has been collected ethically and stored legally, and that any intervention is left to the appropriate authorities (Kennedy et al., 2017; Kumar et al., 2018). Meanwhile from a theoretical point of view, there are many outstanding linguistic and sociological research questions surrounding Internet aggression and how it manifests itself in writing (Pieschl et al., 2015; Waseem et al., 2017).

The question we address here is whether online abusive language is of one type or whether there is discernible variation in the level of abuse found in different subsections of the Internet. We do not claim to have the final answer to this nebulous question, but instead we have addressed one small part of the whole: is the level of abuse found in one Internet domain – namely discussions about English Wikipedia page edits – similar to that found in another domain, that of an online hacking forum?

We show that the type of abusive language occurring in the latter is more closely aligned with the milder levels of abuse of those found in Wikipedia discussions, and consider why this might be. We observe that the online hacking forum tends to contain texts aimed at helping or informing other users, whereas the Wikipedia conversations are inherently more adversarial since they relate to recent page edits and disputes arising. Where abusive language is found in the online hacking forum, it tends to involve profane name-calling, insults and heated disputes, rather than hate speech or threats of violence – those which have tended to be the more prominent causes for public concern.

Note here that we make a distinction between *aggressive* and *offensive* language: the former often involves the latter, but not always so. Offensive language – identifiable word tokens such as swearwords and the like – may offend but is not always used aggressively; sometimes it is used in a jocular fashion, for example. Aggressive language, which more often than not is built on the composition of many words, involves a hostile stance from

[1] Note that this paper quotes texts which many will find offensive and/or upsetting. Please contact the authors if you would prefer to read the article with all quotations removed.

Proceedings of the Second Workshop on Abusive Language Online (ALW2), pages 66–74
Brussels, Belgium, October 31, 2018.

one speaker or writer to another. It is this which might seem to be abusive and which we seek to automatically detect and better understand.

We also distinguish aggressive language from *hate speech* – that which might be characterised as prejudicial diatribes to provoke action, perhaps violent, against a group or groups – and from *cyberbullying* – that which involves a sustained period of persecution against an individual or individuals. Certainly the distinctions are fuzzy at the edges, but these might be thought of as the canonical definitions of these abuse types. We are dealing with what we deem to be one-off instances of aggression in online communities, though if these were shown to be prejudicial against a group, or sustained against an individual, then the instances start to move into hate speech or cyberbullying behaviours.

In both Wikipedia edits and the online hacking forum, abusive comments are infrequent in the community as a whole and the general objective of gaining reputation in the domain dis-incentivises aggressive behaviour. Nevertheless we show that aggressive language which does occur may be detected fairly well by training on the Wikipedia edits corpus – the advantage being that it has been multiply and widely annotated – and setting the threshold for a binary aggression classifier at a fairly moderate level relative to the worst types of abuse found in Wikipedia comments. Future work remains to be done to more broadly characterise intra-community behaviour in different subsections of the Internet.

2 Related work

Offensive language serves many purposes in everyday discourse: from deliberate effect in humour to self-directed profanity to toxic or abusive intent. We are not concerned here with humorous uses of offensive language or with general profanity. Instead we are interested in toxic and abusive behaviour, specifically online harassment involving abusive language, aggression and personal attacks. There has been work on other forms of abusive behaviour, such as hate speech (Warner and Hirschberg, 2012; Kwok and Wang, 2013; Ribeiro et al., 2018) and cyberbullying (Xu et al., 2013; Pieschl et al., 2015), and we put these aside for now as challenging, distinct topics (though with the fuzzy edges described above).

In terms of online harassment, previous work has centred around definitions, automatic detection, and dataset creation – for example the Hate Speech Twitter Annotations and Wikipedia Comments Corpus (Waseem and Hovy, 2016; Wulczyn et al., 2017). Most work has been conducted on English data, with some extensions to other languages (e.g. Arabic (Mubarak et al., 2017), Slovene (Fišer et al., 2017)).

Automated detection approaches have drawn on classic document classification methods for spam detection and sentiment analysis, and tend to use lexical and syntactic features (Nobata et al., 2016; Li et al., 2017; Bourgonje et al., 2018). Machine learning techniques range from logistic regression (Cheng et al., 2015) to support vector machines (Yin et al., 2009) to neural networks (Gambäck and Sikdar, 2017). Our aim here is not especially to push the boundaries on detection techniques – though naturally we wish our classifier to perform fairly well – but rather we are interested in how to make use of existing labelled training data when predicting personal attacks in other corpora.

In case any persuasion is needed that improved understanding, detection and action on abusive language are desirable, there is evidence that experience of online harassment leads to decreased online participation and is connected with oppression, violence and suicide (Dinakar et al., 2011; Sood et al., 2012; Wulczyn et al., 2017). Of course there may be reasons to be concerned about the perpetrator's wellbeing along with that of the victims (Cheng et al., 2017).

3 Training & test corpora

We have an inter-corpus experimental design, in which a document classifier is trained on one dataset and tested on other datasets. Our training data come from the Wikipedia Comments Corpus (WikiComments) (Wulczyn et al., 2017), which contains 115,864 discussion posts extracted from an English Wikipedia dump, judged as personal attacks or harassment by crowdworkers. Ten judgements were collected for each post; hence we have an *attack score* from zero to ten for every post[2], and we assume that the higher the attack score the greater the linguistic aggression shown in writing.

This assumption may be challenged, as we accept that there are many reasons why a text may not be unanimously judged to be an attack or ha-

[2]Note that the original authors scaled the attack score between 0 and 1, whereas we re-scale the scores from 0 to 10.

rassment – properties of the text such as poor grammar which obfuscates meaning, use of slang insults which are not universally known, or sarcastic phrasing which is not interpreted as an attack by all annotators. On the other hand, properties of the annotator, such as fatigue or inattention, inexperience with English or the terminology used, or idiosyncratic linguistic thresholds for attacks and harassment, could all play a part in judgement variation as well. However, over such a large dataset we assume that in terms of aggressive language the texts will be broadly well ordered by their attack scores. Table 1 shows examples randomly drawn from each attack score, zero to ten, along with the number of posts in each class, and the cumulative size of the corpus in reverse order from attack score ten to zero.

The curators of WikiComments used these annotated discussion posts to train a classifier and further label unseen posts in a larger collection of 63 million discussion posts, with a view to large-scale analyses of attacks by unregistered users, moderator actions in response to attacks, and more (Wulczyn et al., 2017). They experimented with different thresholds t where attack scores at or above t would be labelled as attacks, and those below t would not be attacks. They found that the optimal value for t balancing precision and recall was 4.25.

Our intention is to take the texts and attack scores from WikiComments to train a binary aggression classifier for use with other corpora. The question with such a classifier is how to partition the training data for true/false aggression labels: the cut-off could be any attack score value from one to ten. In the following sections we report on classification experiments with each attack score cut-off value and a test corpus sourced from Internet forums.

Our test data come from the CrimeBB Corpus[3], a dataset harvested from several hacking-related websites including HackForums, Antichat and Greysec (Pastrana et al., 2018). The corpus currently contains both English and Russian language data, with plans to incorporate other languages in future. We opted to work only with posts from the HackForums website[4], it being the most popular English language hacking site worldwide.

Among other author intents such as helpfulness, disapproval, sarcasm and gratitude, we manually labelled author aggression as indicated by abusive language in a total of 4123 posts randomly sampled from a selection of HackForums *bulletin boards* (themed discussion pages) from November 2007 to January 2018. All boards are related to hacking (such as 'Cryptography, Encryption, and Decryption', 'Keyloggers', and 'Remote Administration Tools'), as opposed to other interests represented on HackForums such as gaming, entertainment and graphics. Three annotators labelled 2200 posts and agreed to a 'moderate degree' according to Landis & Koch's framework for interpreting Fleiss's kappa (Fleiss, 1971; Landis and Koch, 1977) – i.e. $\kappa = 0.4$ to 0.6. We did not attempt to settle on single annotations for each post, but instead treated all judgements equally, allowing multiple labels both by individual annotators and across different annotators. A single annotator further labelled the remaining 1923 posts.

Posts with aggressive intent are uncommon on HackForums, with only 100 aggressive posts judged to be aggressive by at least one annotator in the total corpus of 4123 posts (2.4%). Note that profane language is more commonly found – which is unsurprising given the casual linguistic register – with 201 posts in this dataset featuring at least one of 'fuck, shit, cunt, jerk, crap, dick' (or derived forms). However, the profanity is often used for humorous purposes, or to defuse potentially confrontational conversations, or simply in a casual way for no purpose at all; it is not always used aggressively (hence the need for manual annotation). This observation underlines the distinction between offensive and aggressive language. Table 2 shows the size of the CrimeBB Corpus, the HackForums subset, and the annotated posts, along with examples of aggressive and non-aggressive posts from HackForums.

4 Classification experiments

We trained a binary aggression classifier on the WikiComments Corpus setting the true/false threshold t at each attack score from 1 to 10 and testing the classifier on our annotated set of 4123 HackForums posts from the CrimeBB Corpus.

4.1 Evaluation

We are interested in the successful classification of aggressive posts only and therefore, rather than reporting precision, recall and F-measures, we re-

[3] Available by application to the Cambridge Cybercrime Centre, https://www.cambridgecybercrime.uk

[4] https://hackforums.net

Attack score	Example	N.posts	Cumulative n.posts
10	fuck u asshole go fuck your own mother fuck fuck fuck fuck a duck fucked a kangaroo you piece of shit go suck your own balls.	3223	3223
9	drop dead. you're a cancer on wikipedia, and i assume you're a cancer on the rest of the world as well. your baby will do better without you.	2842	6065
8	oh right, like "this article sucks and you know it." isn't abusive language. hypocrite. eat shit and die, loser.	2501	8566
7	the link is there. i have an idea. why not delete the page and then re create it a few days later you fucking retard? have you a life at all? is it your life to go around and delete everything here? why not go outside and build a snowman you loser?	2195	10,761
6	next time you remove a section without first following procedures, i will have you permanently barred from wikipedia. i wrote the section and it is completely accurate an unbiased as reported. you are a fool and a vandal and i will be closely monitoring your activities on anything i write.	2706	13,467
5	so this looks like a vendetta. it is stupid. you are pissing me off. this is not what we are here for.	2263	15,730
4	everybody!!! everybody!!!! *redacted* is the undigested part of a giant squid stuck in a sperm whales intestines and testicles	4133	19,863
3	hello, hoaxer. thank you for posting on my talk page. it appears that you enjoy wasting hours first hoaxing and then arguing about it with wikipedia editors on discussion and user pages. all one needs to is track your ip army to see that you are the hoaxer. nobody is falling for your nonsense, especially when you don't sign your posts.	6280	26,143
2	i am aware that most bible thumping christians want to burn this guy alive. i find your assessment far from neutral, i will agf here, but your tone is vitriolic.	9408	35,551
1	the new title doesn't convey what i wanted the section to be about, think of title that conveys the question not just the general subject matter.	22,548	58,099
0	in a legal brief, one might well exclude trial court opinions. in an encyclopedia article, it's a different story, especially when the trial court opinion predates the appellate decision by decades.	57,765	115,864

Table 1: Examples, the number of posts, and the cumulative size (in reverse order) for each attack score subset of the Wikipedia Comments Corpus (Wulczyn et al., 2017).

Corpus	Example	N.posts
CrimeBB		57,733,219
HackForums		40,152,443
Annotated dataset		4123
Non-aggressive	my bet would be install linux and then use spoofing via that	4023
Aggressive	kill yourself. most retarded advice you could give him	100

Table 2: Examples and the number of posts in subsets of the CrimeBB Corpus (Pastrana et al., 2018).

port accuracy as in equation (1):

$$\text{Accuracy} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (1)$$

4.2 Method

All test and training texts were lower-cased and transformed into document-term matrices using the `text2vec` package for R (Selivanov and Wang, 2017). For each value of threshold t from 1 to 10, the training texts were assigned true and false labels according to their attack score s where aggression is true if $s \geq t$.

We trained an extreme gradient boosting (XGBoost) classifier with the R package `xgboost` (Chen et al., 2018). Boosting is an additive technique whereby new models are added to correct the errors made by existing models thus far: models are added sequentially until no further improvements can be made. In gradient boosting, new models predict the residuals or errors of prior models using a gradient descent algorithm. XGBoost is known to work well with sparse matrices, which is the kind of input associated with textual data, and in NLP terms has been shown to perform competitively in sentiment analysis shared tasks (Nasim, 2017; Jabreel and Moreno, 2018).

To avoid over-fitting we set parameters fairly conservatively, with a maximum tree depth of 6, the number of rounds at 10 and early stopping set to 5, gamma at 1, and the learning rate at 0.3. We report classifier accuracy according to equation (1) on gold aggression:true labels in our CrimeBB test corpus. Recall that we do not compare XGBoost with other classifiers, as our focus is on the training data rather than performance. In future work we can investigate other models including neural networks, though logistic regression has in some cases out-performed neural nets in the detection of abusive language (Park and Fung, 2017).

As the value of t increases the size of the aggression:true dataset decreases, as seen in Table 1. To ensure any change in accuracy is not due to the decrease in aggression:true training instances, we run a second experiment in which for all values of t both label subsets (aggression:true and aggression:false) are randomly reduced to 3223 instances – the size of the smallest attack score subcorpus (per the cumulative n.posts column in Table 1). For this latter experiment we report accuracies averaged over one hundred runs to smooth variation in the random sampling process (identified as 'Acc.Control' in Table 3).

4.3 Results

Classification accuracies are shown in Table 3 [5]. It is apparent that in both training data settings – controlled and non-controlled ('all') – the accuracy of aggression identification reduces as the true/false cut-off threshold t increases. In the case of the controlled training data setting there is at first a small increase in accuracy as t rises from 1 to 3. This result suggests that the levels in the Wiki-Comments Corpus most closely matching the aggressive posts on HackForums are those in the attack score range 1 to 5, and that the optimal value of t is between 2 and 3.

To illustrate the rise and fall in classification accuracy as t increases, we plot accuracies as boxplots for the 100 runs in the controlled training data setting (Figure 4.3). The boxplots show medians (the thick horizontal bars), first and third quar-

[5] For comparison with the classifiers trained by Wulczyn et al (2017) we also calculated AUC (area under the curve) measures in the 'all' condition. Our best AUC was .739 with t at 2; Wulczyn et al's best model was a multi-layered perceptron estimating empirical distributions based on character n-grams and this achieved an AUC of .966.

t	N.True posts	Acc. All	Acc. Control
1	58,099	.80	.76
2	35,551	.63	.77
3	26,143	.54	.78
4	19,863	.41	.75
5	15,730	.35	.72
6	13,467	.32	.70
7	10,761	.27	.64
8	8566	.22	.60
9	6065	.19	.52
10	3223	.09	.42

Table 3: Classification accuracy for aggressive posts in the CrimeBB Corpus, with a varying true/false training threshold t from 1 to 10, the size of the aggression:true set in WikiComments for different values of t, accuracy for all training WikiComments instances, and a controlled experiment sampling 3223 true and false instances (averaged over 100 runs).

tiles (Q1, Q3, shown by the hinges), and whiskers extending as far as $1.5 * \text{IQR}$ where IQR is the inter-quartile range between Q1 and Q3. Datapoints beyond the whiskers are outliers and are plotted individually.

4.4 Discussion

It is evident from our classification experiments that levels of linguistic aggression in HackForums tend to be milder than those in WikiComments, if we take the optimal value of t to lie between 2 and 3 (Table 3) whereas for WikiComments it was found to be 4.25 (Wulczyn et al., 2017). A possible explanation for this finding may be the difference in purposes of the two sources for our test and training data: discussion of Wikipedia page edits often end up as arguments between contributors. The fact these arguments may become aggressive or personally offensive at times is unsurprising.

In HackForums, where our test data came from, users often have the intention of educating others, learning from others, buying and selling products, and in many cases discouraging others from acting illegally online (those with a so-called 'white hat' hacking ethos – hackers who identify security vulnerabilities and report them rather than exploit them). HackForums is not an oasis of calm, positive behaviour, however – on the contrary, users can often be off-hand in their comments, dismissive of 'noobs' and 'skids' (script kiddies – a novice or tinkerer), sarcastic and rude. These attitudes, where they do not cross the line into aggressive behaviour, map to our negative label for author intent. Debates about hacking techniques, authorship of code, and user behaviour (e.g. spam, posting out-of-date tutorials, offering hacking tools which don't work as advertised) are frequent. But on the whole, the forum exists for information and technology exchange and the white hat hackers, along with active administrators and a reputation scoring system, help to constrain user behaviour.

Indeed this highly active reputation scoring system may deter aggressive online harassment and allow for users to engender trust in what could otherwise be quite untrustworthy environments (Holt et al., 2016; Décary-Hétu and Leppänen, 2016). Furthermore, online deviant communities such as these tend to be rather homogeneous, particularly involving young males (Hutchings and Chua, 2017). Therefore the targets for any harassment may be off, rather than on, the forum.

Aside from aggression, we also labelled positive texts (which answer others' questions, contain laughter-related word tokens or emoticons, or praise the work of others), neutral texts, and negative texts (including users stating that others cannot or should not do something, sarcasm and arguments). These intent types are the majority labels in our 4123 post subset, with 1562 positive, 2566 neutral and 788 negative occurrences (the posts could be multiply labelled, hence these counts sum to more than 4123). Minority labels are aggression (n=100), users posting to moderate discussion (n=119), and requests to continue discussion in private messaging (n=238).

We further subdivide our set of 100 aggressive forum posts into seven classes: simply aggressive, personal denigration, alludes to violence, refers to disability, features misogyny, homophobia, racism. Personal denigration typically involves name-calling – dismissing someone as an idiot or moron, doubting their technical skills, and so on. The other classes indicate that the author of the post alludes to violence ("I'll cut your neck"), disability ("you're a retard"), misogyny ("stop bitching"), homophobia ("that's gay"), and racism ("fucking jew"). Note that, with the exception of 'simply aggressive' which tends to be

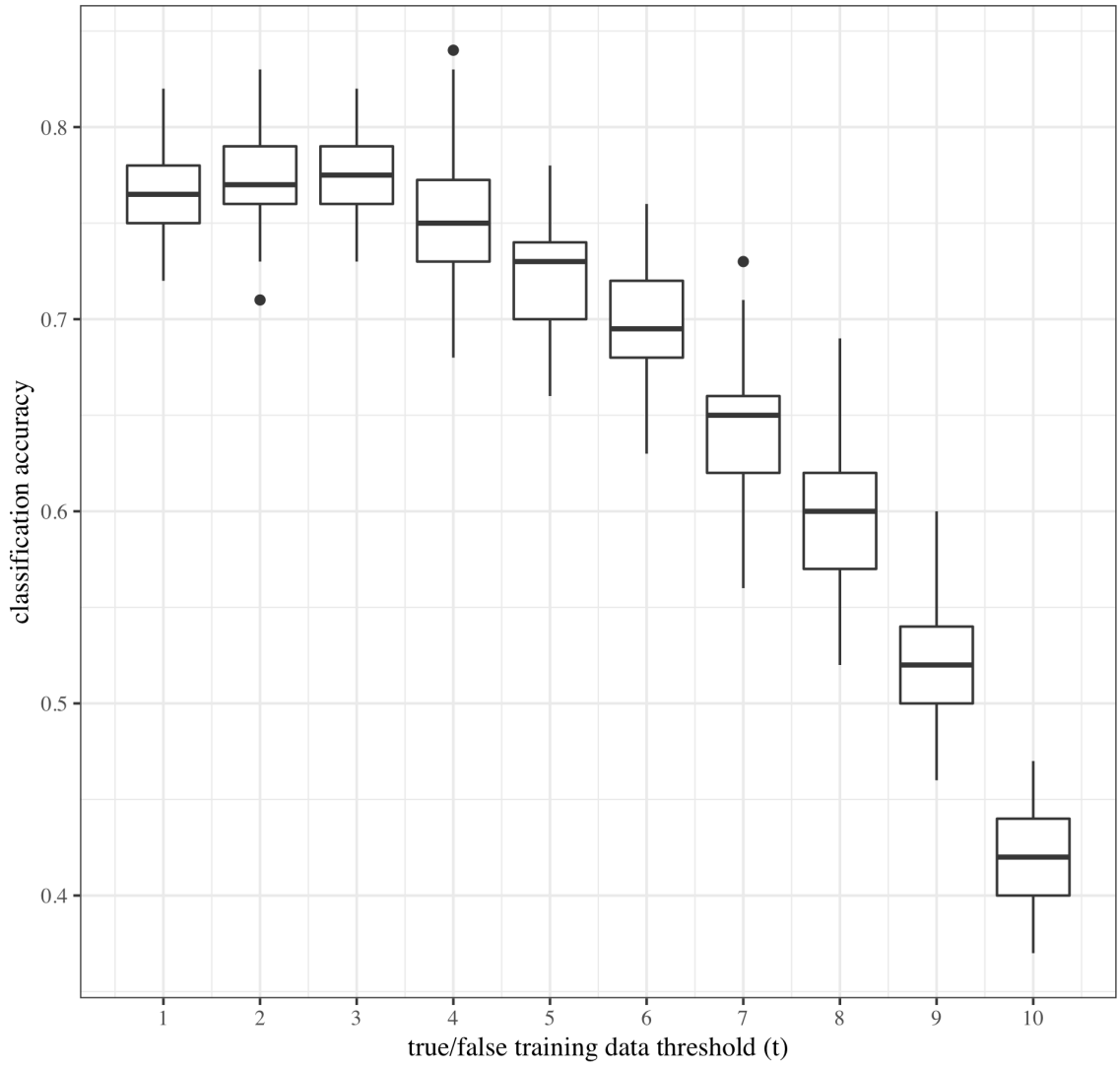

Figure 1: Classifying aggressive posts in the CrimeBB Corpus using controlled training data sizes; with the true/false training threshold t on the x-axis and accuracy on the y-axis, and each data point being 1 of 100 runs randomly sampling the training data.

a fallback if the post falls into no other class, the posts may be assigned multiple labels and that a single annotator undertook labelling. Label counts are shown in Table 4.

We find that most aggressive posts are just that – simply aggressive manners of writing which would be out of place in polite discourse. For example, authors add emphasis with the f-word, including formulaic phrases in acronym form ('gtfo', 'wtf', 'stfu'). The next most common aggression type is personal denigration: most often calling the addressee's intelligence into question, or doubting their motives. After that, the minority labels are those which might feature in hate speech: discriminating against women, homosexuals and ethnicities. In addition, the 'refers to dis-

Label	**Count**
Simply aggressive	48
Personal denigration	37
Refers to disability	7
Includes misogyny	4
Alludes to violence	2
Includes homophobia	1
Includes racism	1

Table 4: Aggression subclass counts in 100 HackForums posts with aggressive intent from the CrimeBB Corpus.

ability' label always involves the words 'retard' and 'retarded' in this 100 post sample. Finally, direct threats of violence are very rare, with only two examples found in this subcorpus.

5 Conclusions & Future work

We have shown that abusive language in an online hacking forum is relatively mild compared to that found in Wikipedia page edit comments. We propose that the tendency of forum users to on the whole engage in constructive and informative discourse results in positive behaviour and non-toxic language. WikiComments, on the other hand, is made up of debates about the rights and wrongs of page edits, and perhaps inevitably this adversarial set up allows more aggressive behaviours to manifest themselves in writing.

In future work we evidently need to annotate more data so that we have more than 100 examples of abusive language from CrimeBB. Due to the low hit rate for abusive language in CrimeBB texts (100 in 4123, for instance) we can investigate automatic annotation of further chunks of the data, along with supervised sampling from those new annotations to check their quality. These labelled data on a larger scale will allow us to analyse more general patterns of behaviour such as individual and community-wide trends over time, how aggression surfaces and is dealt with by moderators, and linguistic facets of aggressive behaviour such as homophobia, racism, misogyny and so on.

We can also investigate other Internet domains such as social media, other forums and potentially the Dark Web, but also other sections of CrimeBB, such as the reputation voting area within HackForums in which we might expect to find more vitriolic interactions given that votes can be both positive and negative and accompanied by review-like texts. Finally, we are also interested in applications of our research, including the questions of desired accuracy of any deployed system, the appropriate actions to take, and the ethics of data collection, analysis and intervention (Kennedy et al., 2017; Thomas et al., 2017). One option could be to create an alert system for forum moderators, thereby offering real-world impact for our work while allowing the appropriate authorities to take action when necessary (Kumar et al., 2018).

Acknowledgments

This work was supported by The Alan Turing Institute's Defence & Security Programme, and the U.K. Engineering & Physical Sciences Research Council. We thank Emma Lenton, Dr Alastair Beresford, and the anonymous reviewers for their support and advice.

References

Peter Bourgonje, Julian Moreno-Schneider, Ankit Srivastava, and Georg Rehm. 2018. Automatic classification of abusive language and personal attacks in various forms of online communication. In *Language Technologies for the Challenges of the Digital Age*. Springer International Publishing.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. 2018. *xgboost: Extreme Gradient Boosting*. R package version 0.6.4.1.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *The 9th International AAAI Conference on Web and Social Media (ICWSM)*.

David Décary-Hétu and Anna Leppänen. 2016. Criminals and signals: An assessment of criminal performance in the carding underworld. *Security Journal*, 29:442–460.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*.

Joseph Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*.

Thomas Holt, Olga Smirnova, and Alice Hutchings. 2016. Examining signals of trust in criminal markets online. *Journal of Cybersecurity*, 2:137–145.

Alice Hutchings and Yi Ting Chua. 2017. Gendering cybercrime. In T. J. Holt, editor, *Cybercrime through an Interdisciplinary Lens*. Oxford: Routledge.

Mohammed Jabreel and Antonio Moreno. 2018. EiTAKA at SemEval-2018 Task 1: An ensemble of n-channels ConvNet and XGboost regressors for emotion analysis of tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*.

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the First Workshop on Abusive Language Online*.

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Tai Ching Li, Joobin Gharibshah, Evangelos E. Papalexakis, and Michalis Faloutsos. 2017. TrollSpot: Detecting misbehavior in commenting platforms. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*.

Zarmeen Nasim. 2017. IBA-Sys at SemEval-2017 Task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

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*.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online*.

Sergio Pastrana, Daniel Thomas, Alice Hutchings, and Richard Clayton. 2018. Crimebb: Enabling cybercrime research on underground forums at scale. In *Proceedings of the 27th International Conference on World Wide Web (WWW'18)*.

Stephanie Pieschl, Christina Kuhlmann, and Torsten Porsch. 2015. Beware of publicity! perceived distress of negative cyber incidents and implications for defining cyberbullying. *Journal of School Violence*, 14:111–132.

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, and Wagner Meira J VirgíÌĄlio A. F. Almeid and. 2018. "like sheep among wolves": Characterizing hateful users on Twitter. In *Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*.

Dmitriy Selivanov and Qing Wang. 2017. *text2vec: Modern Text Mining Framework for R*. R package version 0.5.0.

Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63:270–285.

Daniel Thomas, Sergio Pastrana, Alice Hutchings, Richard Clayton, and Alastair Beresford. 2017. Ethical issues in research using datasets of illicit origin. In *Proceedings of the ACM Internet Measurement Conference (IMC'17)*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*.

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*.

Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. 2013. An examination of regret in bullying tweets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on Web 2.0. In *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*.