2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Generalisation of Cyberbullying Detection

Marc-André Larochelle
Computer Science and Software Engineering Department
Université Laval
Québec, Canada
marc-andre.larochelle.1@ulaval.ca

Richard Khoury
Computer Science and Software Engineering Department
Université Laval
Québec, Canada
richard.khoury@ift.ulaval.ca

***Abstract*—Filtering out Cyberbullying of communities has proven to be a challenge, and efforts have led to the creation of many different datasets to train classifiers. Through these datasets, we will explore the variety of definitions of cyberbullying behaviors and the impact of these differences on the portability of one classifier to another community. We also gain insight on the generalization power of these classifiers. A study of ensemble models combining these classifiers will help us understand how they interact with each other.**

***Index Terms*—natural language processing, deep learning, cyberbullying, cross-domain generalisation**

I. Introduction

While online social interactions bring together people in an unprecedented way, they have come with a negative impact, the spread of cyberbullying and its social toll. A Canadian study [1] found that 17% of 15-to-29-year-old Internet users experienced cyberbullying, that the problem disproportionately affected women (19%), low-income people (24%), and homosexuals (34%), and that 20% of victims developed emotional, psychological or mental health conditions as a result. Although many solutions to this problem have been proposed, none of them agree on what constitutes cyberbullying. For example, [2] has found that cyberbullying can include general insults, cursing, sexism or hate speech, defamation, sexual harassment, threats or blackmail, and even polite support of the aggressor, directed against a victim or their friends and family, which may occur once or be repeated, which may or may not involve a power imbalance between the aggressor and victim, and which may be an intentional attack or a misunderstanding.

In this paper, we explore the relationship and distinctions that exist between datasets that implement different definitions of cyberbullying, the possibilities and limitations to applying a system trained on one dataset to filter cyberbullying in a new context, and to discover empirically the best way to combine them to create a generalized cyberbullying detection system.

II. Datasets

For our research, we have selected eight datasets pertaining to various types of behaviors that fit the general definition of cyberbullying. Many of these datasets have labels to designate specific types of behaviors which are either not found in other datasets or defined in a different manner. In order to make the datasets directly comparable, we merged all these labels into a positive "cyberbullying" class, opposing the much more common negative class of messages that are not cyberbullying. In addition, some of these datasets came divided into training and testing sets. For the others, we randomly divided the datasets into a 20% test set and 80% training set, which we further divided into 80% training and 20% validation sets.

A. Hate Speech and Offensive Language[1]: This dataset was collected by [3] by searching Twitter for tweets containing hate speech terms from the lexicon *Hatebase.org*. Tweets were randomly selected and annotated by three or more annotator to one of three labels: if it contains hate speech, offensive language without hate speech, or neither. A tweet's final label was the majority decision, and tweets for which no majority decision existed were filtered out. This gave a corpus of 4,163 tweets that do not contain hate speech nor offensive language, 19,190 that contain offensive language, and 1,430 that are considered hate speech, making it the only corpus in our study imbalanced in favor of the positive class.

B. Racism and Sexism[2]: The authors of [4] designed a list of slurs and terms identifying religious, sexual, gender, and ethnic minorities. They then sampled Twitter for tweets using these words. They annotated each sampled tweet as sexist, racist, or neither, and had an expert review it to mitigate possible bias. The dataset is available from [5].

C. Bullying[3]: This dataset was gathered by the authors of [6] from the question-answering (QA) platform *Formspring*. Each QA pair was labeled for cyberbullying and identified which words or phrases were the reason. This created a labeled dataset of 12,773 QA pairs. We follow the work of [5] and concatenate the question and answer into as single message.

D. Insults in social commentary: This Kaggle dataset [7] was gathered from an unspecified social networking site, and its comments were labeled as insulting or not. However, it applies a narrower definition and only labels positive messages if they are explicitly insulting to a specific member of the on-going conversation. Messages that insult public figures, messages that use racial slurs not directed at a specific person,

This research was made possible by the financial, material, and technical support of Two Hat Security Research Corp., and the financial support of the Canadian research funding agency MITACS.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

[1] https://github.com/t-davidson/hate-speech-and-offensive-language
[2] https://github.com/sweta20/Detecting-Cyberbullying-Across-SMPs
[3] www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection

and subtle insults, are all counted as negative-class messages.

E. Hate Speech[4]: This dataset is composed of messages scraped from the white-supremacist internet forum *Stormfront* by [8]. The messages were labeled into one of four classes. The "hate" class is for messages that are (i) deliberate attack (ii) directed towards a specific group of people (iii) and motivated by aspects of the group's identity. This definition is different from traditional hate speech: for example using a racial slur in an offhand manner is not considered a hate message. The "relation" class is for messages that fit in the hate class when read within the context of the conversation in which they appear. The "skip" class contains non-English messages and gibberish. Finally, the "non-hate" class is for messages that do not fit the other three categories. The dataset is composed of 9,507 non-hate messages, 1,196 hate messages, 168 relation messages, and 73 skip messages. We redefined the hate class as our positive class and the others as our negative class. The rationale for including relation in the negative class is that most cyberbullying detection systems work on a line-by-line basis and will mark such messages without context as acceptable.

F. Toxic comment classification: This Kaggle dataset [9] was gathered from Wikipedia talk pages. Each comment was annotated as toxic (21,384 messages), severely toxic (1,962), obscene (12,140), threat (689), insult (11,304), and identity hate (2,117). Unlike the previous datasets, each comment can be tagged with multiple labels. We assign a comment to our positive class if it contains any one or more of these labels.

G. Unintended bias in toxicity classification: This Kaggle dataset [10] was obtained from a news website comment filter system called *Civil Comments*. Its 1,999,514 comments were assigned six labels: severe toxicity, obscene, threat, insult, identity attack, and sexual explicit. All labels chosen by half or more of annotators was applied to the comment. We consider a comment as positive class if any label was applied to it, for a total 1,358,749 positive-class comments.

H. Personal attacks and harassment[5]: The authors of [11] sampled the Wikipedia talk page comments posted between 2004 and 2015, and added a sample of comments from a set of users blocked for violating Wikipedia's policy on personal attacks. Annotators were asked if each comment contains a personal attack or harassment, whether it is targeted at the recipient or a third party, if it is being reported or quoted, and if it is another kind of attack or harassment. This resulted in a corpus of 115,859 comments, of which 13,590 were found to contain personal attack or harassment. We note that there is an overlap between the training set of dataset D and the test set of this dataset.

III. Dataset vocabulary comparison

The eight datasets all pertain to some aspects of cyberbullying [2], although only [6] explicitly names it as such and the problematic behaviors they monitor varies greatly, from the focused scope of the Twitter corpora to the wide range of behaviors labeled in the Kaggle datasets. Some behaviors are common; hate speech is labelled in five of the eight datasets. And some behaviors are unique to some corpora; threats are explicitly noted in only two corpora and sexually-explicit comments in one. This wide variety of forms of cyberbullying and its consequences are the main focus of our study.

To begin, we will study the impact of this diversity on the vocabulary of the datasets. First, we divide each dataset into its positive (+) and negative (-) classes and treat each one separately.To compute cosine similarity, we convert each dataset class into its bag-of-word representation and compute the TFIDF value of each word:

$$tfidf(w, c) = (1 + \log n_{w,c}) \times \log \frac{C}{C_w} \quad (1)$$

where the tfidf value of word w in dataset class c is the log normalization of the number of times the word occurs in the dataset class ($n_{w,c}$) times the inverse log of the number of dataset classes C (16) and C_w the number of datasets classes containing word w. Using the log normalization of the word count (defined as 0 if a word does not occur in a corpus) instead of the word count helps mitigate the impact of the extreme difference in the size of our datasets, by focusing on the order of magnitude of the counts instead of their values. Using the TFIDF-weighted word vectors, we then compute the cosine similarity between each pair our 16 classes, the results are presented in Table I. These results show that, for 6 of the 8 datasets, the most similar dataset to its positive class is its negative class and vice-versa. The exceptions are datasets F and H, the two Wikipedia talk page corpora, which are more similar to each other than to their positive/negative classes.

What is surprising in Table I is how low the similarity values are. Datasets A, B, C, D and E have nearly no similarity to any other dataset. Datasets A and B, both from Twitter, have no more similarity to each other than they do to datasets from other sources. The positive classes of datasets A and E, both focusing on hate speech, have no more in common than they do to other datasets, positive or negative. Even the similarity between the positive and negative class of each dataset is rather low. Normally we would expect language to be homogeneous throughout a dataset and to vary mainly on the presence or absence of class-specific cyberbullying keywords, and since those would be a minority of the words used the similarity between classes should be high. The low values observed indicate the opposite, that the positive and negative classes of each dataset differ also in non-cyberbullying vocabulary.

The low similarity between the positive classes of different datasets shows that the vocabulary marking cyberbullying varies greatly. This can be attributed in part to the different cyberbullying behaviors each dataset measures and to the different platforms each dataset comes from, and to the diversity of the English language. The consequence is that we should expect a system trained to recognize cyberbullying in one dataset to have difficulty picking out cyberbullying in another.

[4]https://github.com/aitor-garcia-p/hate-speech-dataset

[5]https://doi.org/10.6084/m9.figshare.4054689.v6

TABLE I
COSINE SIMILARITY BETWEEN EACH PAIR OF DATASETS

	A_-	A_+	B_-	B_+	C_-	C_+	D_-	D_+	E_-	E_+	F_-	F_+	G_-	G_+	H_-	H_+
A_-	1.000															
A_+	0.212	1.000														
B_-	0.013	0.010	1.000													
B_+	0.009	0.007	0.326	1.000												
C_-	0.014	0.027	0.011	0.006	1.000											
C_+	0.005	0.016	0.003	0.002	0.188	1.000										
D_-	0.014	0.013	0.011	0.009	0.013	0.004	1.000									
D_+	0.013	0.013	0.009	0.007	0.008	0.005	0.249	1.000								
E_-	0.013	0.012	0.015	0.011	0.016	0.003	0.024	0.013	1.000							
E_+	0.009	0.010	0.010	0.008	0.008	0.002	0.015	0.016	0.094	1.000						
F_-	0.026	0.024	0.032	0.021	0.042	0.010	0.048	0.023	0.074	0.032	1.000					
F_+	0.024	0.029	0.033	0.024	0.041	0.018	0.048	0.037	0.066	0.038	0.274	1.000				
G_-	0.030	0.030	0.035	0.022	0.048	0.012	0.052	0.025	0.067	0.032	0.284	0.151	1.000			
G_+	0.036	0.036	0.046	0.034	0.051	0.013	0.074	0.042	0.085	0.049	0.276	0.206	0.545	1.000		
H_-	0.026	0.024	0.036	0.024	0.042	0.010	0.051	0.024	0.083	0.038	0.519	0.301	0.258	0.283	1.000	
H_+	0.021	0.024	0.031	0.022	0.035	0.015	0.042	0.032	0.062	0.036	0.228	0.510	0.126	0.182	0.303	1.000

IV. GENERALIZATION EXPERIMENT

To begin, we study the generalization of a cyberbullying detector trained on one of the datasets of Section II to the other datasets. This is an important result to document: while every dataset has trained detectors that perform well on a test corpus subset of itself, experiments on other datasets are rare, and when done [2], [5], [12] they do not explore the limits of this transfer. Moreover, real-world online communities are very heterogeneous in style, content, and acceptability standards, and so establishing the portability of a cyberbullying detector trained for one community is very important.

A. Model and Training

Each message of a dataset is lowercased and tokenized. We pad the messages to match the longest message of a batch instead of truncating or padding the entire dataset, which would result in either memory waste or lost information. We use FastText pre-trained on Common Crawl data featuring 300 dimensions and 2 million word vectors with subword information[6] to convert the words into vector representations, of which we concatenate a 60-dimensional binary vector of common characters; each character appearing in a word is a 1 in this vector. This makes the system robust to misspellings and typos: a misspelled word may be a distant vector in word embedding space, but will be nearby in character space [13].

For our experiments, getting state-of-the-art results is not as important as getting comparable results on all datasets. We thus opted for a model that features two bi-LSTM layers of 128 hidden units each, followed by a scaled-dot product attention, global max and average pooling, and finally three linear layers the size of the concatenated pooling and attention layers. We use layer normalization on the input, after the two bi-LSTM layers, and after the attention layer. We perform dropout after the initial layer normalization and before the last linear layer. The activation function is the GELUs [14]. This architecture gave good results in previous works [5] and in Kaggle competitions [9], [10]. We use a learning rate of 0.05, a decaying gamma of 0.6 every epoch and cross-entropy loss. We use a batch size of 128 messages and train on each dataset for 15 epochs using the Fused Adam optimiser from Nvidia's Apex library and mixed-precision [15] to reduce training time

[6]https://github.com/facebookresearch/fastText

B. Results and Analysis

Table II presents the precision and recall performance of our model when trained using each of our datasets and tested on another. An ideal filter will block all cyberbullying messages (high recall) and no messages that are not cyberbullying (high precision). The F1-score combines both metrics, but the averaging would masks performance details.

As expected, we find that classifiers trained on each dataset have wildly varying performances on other test datasets, and their precision or recall can drop by as much as 0.8. However, the results are not uniformly bad. The top-performing classifier on each test dataset is not always the one trained on it, and some classifiers perform as well or better on other datasets compared to their own test dataset. This means that some generalization of cyberbullying detection is possible. Interestingly, good candidates for generalization are not related to similar data sources: datasets A and B both come from Twitter and datasets F and H both come from Wikipedia, and while classifier B does perform well on dataset A, the other three show no special affinity for their similar datasets. Likewise, modeling a similar behavior does not guarantee generalization: datasets A and E both focus on hate speech, yet they each have very poor recall on the other's test set.

We can see that every model achieves an average precision between 0.45 and 0.63, meaning half the messages each one labels as positive class actually belongs to the negative class. This is a direct consequence of the problem described in section III: a lot of non-cyberbullying words are observed only in the positive or negative class of a dataset. When nearby words in word vector space are observed in an opposite-class message in a test dataset, it causes the test message to be misclassified. Looking next at recall results, we can see that

classifiers often perform worse on other datasets. Classifiers trained on datasets B, C, D, and E pick out less than half the positive messages on average across datasets. These are also our four smallest datasets. The limited variety of messages they have been trained on, compounded by the differences between datasets shown in section III, means filters trained on these datasets cannot generalize to new situations. By contrast, datasets F and G achieve the highest recall scores. These are also the two largest datasets, and the two that label the largest range of cyberbullying behaviors. Having seen a greater variety of cyberbullying messages has allowed them to generalize better to new datasets.

Looking at performance by dataset, we find that all models achieve their highest precision on dataset A, and all but two achieving better than 0.5 recall. This is because it is the only dataset imbalanced in favour of the positive class. This makes the classification task easier; classifiers can be less discriminating without mislabeling negative-class messages. Every classifier achieves a significantly lower precision on every other dataset, including the one it is trained for, indicating that all classifiers routinely mislabel negative-class messages as positive class. On the other hand, models achieve most of the lowest precision and recall scores on dataset E. Aside from the model trained specifically on dataset E, every model fails here, with on average only one-third of labeled messages being actually positive-class and one-sixth of positive-class messages being identified as such. Dataset E, like dataset D, includes a clear intent to attack in the definition of its positive class, and thus has messages with off-hand racial slurs labeled as negative-class. But unlike dataset D, dataset E was collected from a community where off-hand hateful messages are commonplace. These negative-class messages correspond to positive-class messages in other communities and cause a low precision. The low recall is due to the fact that many of its positive-class messages use racist imagery and idioms (e.g. "African blood") rather than explicit hate-speech vocabulary.

Finally, there is a lot of variability in the performances of the classifiers trained on each dataset. For instance, classifiers B and D have some of the worst precision and recall scores, but also sometimes outperform other classifiers. Likewise, classifiers F, G and H get some of the best performances but also occasionally drop dramatically. This further highlights the unreliability and unpredictability of transferring a pre-trained cyberbullying classifier to a new community.

TABLE II
CROSS-DATASET PRECISION AND RECALL

		Test dataset							
		A	B	C	D	E	F	G	H
Precision	A	0.98	0.57	0.31	0.53	0.27	0.56	0.48	0.77
	B	0.96	**0.77**	0.27	0.35	0.22	0.44	0.23	0.43
	C	**0.99**	0.69	**0.53**	0.63	**0.66**	**0.65**	0.47	**0.87**
	D	0.97	0.45	0.19	**0.75**	0.19	0.59	0.38	0.77
	E	0.97	0.72	0.41	0.48	0.55	0.49	0.24	0.64
	F	0.95	0.51	0.21	0.51	0.36	0.51	0.62	0.77
	G	0.97	0.52	0.30	0.55	0.37	0.58	**0.80**	0.81
	H	0.98	0.43	0.26	0.56	0.23	0.61	0.68	0.84
Recall	A	**0.98**	0.21	0.65	0.67	0.17	0.65	0.14	0.63
	B	0.53	**0.78**	0.25	0.11	0.02	0.11	0.04	0.11
	C	0.63	0.17	0.57	0.68	0.06	0.47	0.12	0.47
	D	0.44	0.10	0.68	0.70	0.10	0.48	0.20	0.53
	E	0.15	0.07	0.28	0.24	**0.60**	0.19	0.24	0.17
	F	0.93	0.30	**0.84**	**0.86**	0.35	**0.93**	0.49	**0.86**
	G	0.81	0.25	0.74	0.79	0.37	0.89	**0.57**	0.76
	H	0.75	0.15	0.78	0.82	0.11	0.79	0.30	0.77

V. ENSEMBLE EXPERIMENTS

Next, we study if it is possible to combine a set of classifiers, each trained on a different corpus, into an accurate general classifier. To explore this question, we compare together different ensemble architectures built from our individual models.

A. Ensemble Models

We implemented five ensemble models that combine our existing classifiers in different ways without retraining them.

1. Linear layer (LL): The outputs of the individual classifiers (after softmax) are combined in a linear layer. This layer is trained using the same hyperparameters as the models, and dropout ensures it does not overfit to a single model's decision.

2. Democratic voting (DV): Each classifier casts a vote based on its classification of a message, and the winning class is simply the one with the most votes.

3. Sum voting (SV): Instead of the all-or-nothing vote for one class of DV, each classifier votes for both the negative and positive class with the probability it assigns to each class (after softmax). A classifier confident in its result will cast a strong vote for one class while another will cast almost equal votes for both classes; however, several weak votes in one class may still overrule a single strong vote in the other.

4. Maximum wins (MW): The classifier with the maximum confidence in its output picks the message's class.

5. Thresholding: If any classifier identifies a message as positive class with a confidence above a threshold, it is labeled as such regardless of the output of the others. We implemented two variations of this classifier, one with the threshold at 0.5 (T0.5), the lowest possible confidence to assign a message to the positive class, and the other at 0.95 (T0.95).

6. Dataset merger (DM): As a baseline, we merged together all datasets and trained a new classifier on it.

B. Results and Analysis

Table III gives the precision and recall value of each ensemble technique when applied to each test dataset. Compared to Table II, we can see the ensemble models have generally better precision and worse recall. The F1-scores, which can be computed from these values, of the ensemble models are comparable or better to those of the individual classifiers. This means that combining the information learned from training on different datasets improves the overall performance.

Looking at individual ensembles, we can see that DV, SV and MW have similar behaviors: they achieve some of the best precision scores and worst recall scores of all systems. This indicates that most classifiers mislabel most positive-class messages. As a result, when the ensemble decides a message belongs in the positive class, it is usually right. However, most

positive-class messages are only recognized by a minority or by low-confidence classifiers, and thus recall is low.

The LL ensemble achieves slightly worse precision but much better recall than DV, SV and MW. By learning to combine the individual classifiers, it catches a lot more positive-class messages, but mislabels a few more negative-class messages. In addition, LL actually outperforms all but one of the individual classifiers in precision and in recall and all of them in F1-score, again confirming the benefit of learning a combination of classifiers.

The T0.5 model achieves the top recall and lowest precision scores by wide margins. This means that most positive-class messages are weakly labeled as such by at least one classifier. By increasing the threshold, T0.95 improves its precision in all tests as weakly-positive messages are no longer marked in the positive class. However, the recall value decreases sharply as well, indicating that there are a lot of positive-class messages that not a single classifier can confidently recognize.

The DM approach gives the best overall performance and F1-score. It surpasses LL and T0.95 in precision and is second only to T0.5 in recall, and it is nearly as good as DV, SV and MW in precision with much better recall. It thus seems that merging datasets in a single classifier is better than combining multiple individual classifiers. This empirical conclusion stems also from our earlier analysis in Section IV, where a diversity of language use and of cyberbullying behaviors was key to achieving good results. The DM classifier is trained on the largest possible vocabulary and the largest set of different behaviors. Moreover, this merger will reduce the problem of having neutral-meaning words that appear only in one class of one dataset and the opposite class of another and thus confuse the classification. After the merger, these words appear in both classes and no longer have strongly influence the classification.

TABLE III
ENSEMBLE MODELS PRECISION AND RECALL

		Test dataset							
		A	B	C	D	E	F	G	H
Precision	LL	0.97	0.56	0.28	0.55	0.38	0.59	0.82	0.82
	DV	**0.99**	0.63	**0.45**	**0.68**	0.30	**0.78**	0.87	**0.93**
	SV	0.99	0.57	0.40	0.62	0.41	0.75	**0.89**	0.93
	MW	0.99	0.65	0.35	0.60	**0.46**	0.72	0.89	0.91
	T0.5	0.93	0.62	0.14	0.44	0.31	0.40	0.33	0.58
	T0.95	0.98	0.70	0.32	0.57	0.37	0.60	0.82	0.82
	DM	0.98	**0.72**	0.35	0.63	0.40	0.58	0.81	0.83
Recall	LL	0.87	0.30	0.78	0.82	0.29	0.91	0.50	0.78
	DV	0.69	0.14	0.66	0.69	0.07	0.52	0.11	0.52
	SV	0.78	0.17	0.69	0.74	0.12	0.63	0.16	0.62
	MW	0.87	0.27	0.65	0.67	0.11	0.58	0.163	0.57
	T0.5	**0.99**	**0.83**	**0.91**	**0.93**	**0.72**	**0.98**	**0.75**	**0.91**
	T0.95	0.96	0.47	0.78	0.78	0.18	0.84	0.32	0.75
	DM	0.98	0.74	0.65	0.77	0.38	0.89	0.54	0.82

VI. CONCLUSION

In this paper, we conducted an in-depth study of the relationship between eight cyberbullying datasets and the systems that can be trained from them. First, we studied the datasets and what they tell us about cyberbullying behaviors. Next we studied the similarity in vocabulary between them. We then trained deep neural networks and studied how they can be transferred from one domain to another. Finally, we studied approaches for combining the classifiers into ensemble models.

Our paper has highlighted four major conclusions. First, there is little agreement on the definition of cyberbullying, the behaviors that comprise it, or how to measure and label them. Our second conclusion is that there is very little language in common between datasets and many neutral-meaning words are labeled in contradictory ways, which makes transferring systems difficult. Our third conclusion is that the condition to facilitate transferability is to have a system trained on as diverse a dataset as possible, both in terms of language use and in terms of behaviors labeled. Finally, if one wishes to combine the knowledge from different datasets, the best way of doing this is to merge the datasets and train a single system.

REFERENCES

[1] D. W. Hango, *Cyberbullying and cyberstalking among Internet users aged 15 to 29 in Canada*. Statistics Canada Ottawa, Ontario, 2016.

[2] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," *PloS one*, vol. 13, no. 10, 2018.

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ser. ICWSM '17, 2017, pp. 512–515.

[4] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 88–93.

[5] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Advances in Information Retrieval*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham: Springer International Publishing, 2018, pp. 141–153.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine Learning and Applications and Workshops*, vol. 2, 2011, pp. 241–244.

[7] "Detecting Insults in Social Commentary." [Online]. Available: https://kaggle.com/c/detecting-insults-in-social-commentary

[8] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate Speech Dataset from a White Supremacy Forum," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 11–20. [Online]. Available: https://www.aclweb.org/anthology/W18-5102

[9] "Toxic Comment Classification Challenge." [Online]. Available: https://kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[10] "Jigsaw Unintended Bias in Toxicity Classification." [Online]. Available: https://kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[11] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17. International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399. [Online]. Available: http://doi.org/10.1145/3038912.3052591

[12] C. Emmery, B. Verhoeven, G. De Pauw, G. Jacobs, C. Van Hee, E. Lefever, B. Desmet, V. Hoste, and W. Daelemans, "Current limitations in cyberbullying detection: on evaluation criteria, reproducibility, and data scarcity," *arXiv preprint arXiv:1910.11922*, 2019.

[13] E. Brassard-Gourdeau and R. Khoury, "Subversive toxicity detection using sentiment information," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 1–10.

[14] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2018. [Online]. Available: http://arxiv.org/abs/1606.08415

[15] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.