Lang Resources & Evaluation (2021) 55:597–633
https://doi.org/10.1007/s10579-020-09509-1

ORIGINAL PAPER

Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity

Chris Emmery[1,2] · Ben Verhoeven[2] · Guy De Pauw[2] · Gilles Jacobs[3] · Cynthia Van Hee[3] · Els Lefever[3] · Bart Desmet[3] · Véronique Hoste[3] · Walter Daelemans[2]

Accepted: 24 September 2020 / Published online: 16 November 2020

Abstract The detection of online cyberbullying has seen an increase in societal importance, popularity in research, and available open data. Nevertheless, while computational power and affordability of resources continue to increase, the access restrictions on high-quality data limit the applicability of state-of-the-art techniques. Consequently, much of the recent research uses small, heterogeneous datasets, without a thorough evaluation of applicability. In this paper, we further illustrate these issues, as we (i) evaluate many publicly available resources for this task and

The work presented in this article was carried out in the framework of the AMiCA (IWT SBO-project 120007) project, funded by the government agency for Innovation by Science and Technology (IWT).

✉ Chris Emmery
cmry@pm.me

Ben Verhoeven
ben.verhoeven@uantwerpen.be

Guy De Pauw
guy.depauw@uantwerpen.be

Gilles Jacobs
gilles.jacobs@ugent.be

Cynthia Van Hee
cynthia.vanhee@ugent.be

Els Lefever
els.lefever@ugent.be

Bart Desmet
bart.desmet@ugent.be

Véronique Hoste
veronique.hoste@ugent.be

Walter Daelemans
walter.daelemans@uantwerpen.be

demonstrate difficulties with data collection. These predominantly yield small datasets that fail to capture the required complex social dynamics and impede direct comparison of progress. We (ii) conduct an extensive set of experiments that indicate a general lack of cross-domain generalization of classifiers trained on these sources, and openly provide this framework to replicate and extend our evaluation criteria. Finally, we (iii) present an effective crowdsourcing method: simulating real-life bullying scenarios in a lab setting generates plausible data that can be effectively used to enrich real data. This largely circumvents the restrictions on data that can be collected, and increases classifier performance. We believe these contributions can aid in improving the empirical practices of future research in the field.

Keywords Cyberbullying detection · Cross-domain evaluation · Reproducibility · Crowdsourcing · Data enrichment

1 Introduction

Learning to accurately classify rare phenomena within large feeds of data poses challenges for numerous applications of machine learning. The volume of data required for representative instances to be included is often resource-consuming, and limited access to such instances can severely impact the reliability of predictions. These limitations are particularly prevalent in applications dealing with sensitive social phenomena such as those found in the field of forensics: e.g., predicting acts of terrorism, detecting fraud, or uncovering sexually transgressive behavior. Their events are complex and require rich representations for effective detection. Conversely, online text, images, and meta-data capturing such interactions have commercial value for the platforms they are hosted on and are often off-limits to protect users' privacy.

An application affected by such limitations with increasing societal importance and growing interest over the last decade is that of cyberbullying detection. Not only is it sensitive, but the data is also inherently scarce in terms of public access. Most cyberbullying events are off-limits to the majority of researches, as they take place in private conversations. Fully capturing the social dynamics and complexity of these events requires much richer data than available to the research community up until now. Related to this, various issues with the operationalization of cyberbullying detection research were recently demonstrated by Rosa et al. (2019), who share much of the same concerns as we will discuss in this work. While their work focuses on methodological rigor in prior research, we will focus on the core limitations of the domain and complexity of cyberbullying detection. Through an

[1] CSAI, Tilburg University, Tilburg, The Netherlands

[2] CLiPS, University of Antwerp, Antwerp, Belgium

[3] LT3, Ghent University, Ghent, Belgium

evaluation of the current advances on the task, we illustrate how the mentioned issues affect current research, particularly cross-domain. Finally, we demonstrate crowdsourcing in an experimental setting to potentially alleviate the task's data scarcity. First, however, we introduce the theoretical framing of cyberbullying and the task of automatically detecting such events.

1.1 Cyberbullying

Asynchrony and optional anonymity are characteristic of online communication as we know it today; it heavily relies on the ability to communicate with people who are not physically present, and stimulates interaction with people outside of one's group of close friends through social networks (Madden et al. 2013). The rise of these networks brought various advantages to adolescents: studies show positive relationships between online communication and social connectedness (Bessiere et al. 2008; Valkenburg and Peter 2007), and that self-disclosure on these networks benefits the quality of existing and newly developed relationships (Steijn and Schouten 2013). The popularity of social networks and instant messaging among children has them connecting to the Internet from increasingly younger ages (Ólafsson et al. 2013), with 95% of teens[1] ages 12–17 online,of which 80% are on social media (Lenhart et al. 2011). For them, however, the transition from social interaction predominantly taking place on the playground to being mediated through mobile devices (Livingstone et al. 2011) has also moved negative communication to a platform where indirect and anonymous interaction has a window into homes.

A range of studies conducted by the Pew Research Center[2], most notably (Lenhart et al. 2011), provides detailed insight into these developments. While 78% of teens report positive outcomes from their social media interactions, 41% have experienced at least some adverse outcomes, ranging from arguments, trouble with school and parents, physical fights and ending friendships. From 19% bullied in the 12 months prior to the study, 8% of all teens reported this was some form of cyberbullying. These numbers are comparable to other research (Robers et al. 2015; Kann et al. 2014) (7% for Grades 6–12, and 15% Grades 9–12 respectively). Bullying has for a while been regarded as a public health risk by numerous authorities (Xu et al. 2012), with depression, anxiety, low self-esteem, school absence, lower grades, and risk of self-medication as primary concerns.

The act of cyberbullying—other than being conducted online—shares the characteristics of traditional bullying: a power imbalance between the bully and victim (Sharp and Smith 2002), the harm is intentional, repeated over time, and has a negative psychological effect on the victim (DeHue et al. 2008). With the Internet as a communication platform however, some additional aspects arise: location, time, and physical presence have become an irrelevant factor in the act. Accordingly, several categories unique to this form of bullying are defined (Willard 2007; Beran and Li 2008): *flaming* (sending rude or vulgar messages), *outing* (posting private information or manipulated personal material of an individual without consent),

[1] Survey conducted in 2011 among 799 American teens. Black and Latino families were oversampled.

[2] http://www.pewinternet.org.

harassment (repeatedly sending offensive messages to a single person), *exclusion* (from an online group), *cyberstalking* (terrorizing through sending explicitly threatening and intimidating messages), *denigration* (spreading online gossips), and *impersonation*. Moreover, in addition to optional anonymity hiding the critical figures behind an act of cyberbullying, it could also obfuscate the number of actors (i.e., there might only be one even though it seems there are more). Cyberbullying acts can prove challenging to remove once published; messages or images might persist through sharing and be viewable by many (as is typical for hate pages), or available to a few (in group or direct conversations). Hence, it can be argued that any form of harassment has become more accessible and intrusive. This online nature has an advantage as well: in theory, platforms record these bullying instances. Therefore, an increasing number of researches are interested in the automatic detection (and prevention) of cyberbullying.

1.2 Detection and task complexity

The task of cyberbullying detection can be broadly defined as the use of machine learning techniques to automatically classify text in messages on bullying content, or infer characteristic features based on higher-order information, such as user features or social network attributes. Bullying is most apparent in younger age groups through direct verbal outings (Vaez et al. 2004), and more subtle in older groups, mainly manifested in more complex social dynamics such as exclusion, sabotage, and gossip (Privitera and Campbell 2009). Therefore, the majority of work on the topic focuses on younger age groups, be it deliberately or given that the primary source for data is social media—which will likely result in these being highly present for some media (Duggan 2015). Apart from the well-established challenges that language use poses (e.g., ambiguity, sarcasmdialects, slang, neologisms), two factors in the event add further linguistic complexity, namely that of actor *role* and associated *context*. In contrast to tasks where adequate information is provided in the text of a single message alone, to completely map a cyberbullying event and pinpoint bully and victim implies some understanding of the dynamics between the involved actors and the concurrent textual interpretation of the *register*.

1.3 Register

Firstly, to understand the task of cyberbullying detection as a specific domain of text classification, one should consider the full scope of the register that defines it. The bullying categories discussed in Sect. 1.1 include some initial cues that can be inferred from text alone; flaming being the most obvious through simple curse word usage, slurs, or other profanity. Similarly, threatening or intimidating messages that fall under cyberstalking are clearly denoted by particular word usage. The other categories are more subtle: outing could also be done textually, in the form of a phone number, or pieces of information that are personal or sensitive in nature. Denigration would include words that are not blatantly associated with abusive acts; however, misinformation about sensitive topics might for example be paired with a

victim's name. One could further extend these cues based on the literature (as also captured in Hee et al. (2015)) to include bullying event cues, such as messages that serve to defend the victim, and those in support of the bully. The linguistic task could therefore be framed (partly based on Van Hee et al. (2018)) as *identifying an online message context that includes aggressive or hurtful content against a victim*. Several additional communicative components in these contexts further change the interpretation of these cues, however.

1.4 Roles

Secondly, there is a commonly made distinction between several actors within a cyberbullying event. A naive role allocation includes a bully *B*, a victim *V* and bystander *BY*, the latter of whom may or may not approve of the act of bullying. More nuanced models such as that of Xu et al. (2012) include additional roles (see Fig. 1 for a role interaction visualization), where different roles can be assigned to one person; for example, being bullied and reporting this. Most importantly, all shown roles can be present in the span of one single thread on social media, as demonstrated in Table 1. While some roles clearly show from frequent interaction with either a positive or negative sentiment (*B*, *V*, *A*), others might not be observable through any form of conversation (*R*, *BY*), prove too subtle, or not distinguishable from other roles.

1.5 Context

Thirdly, the content of the messages has to be interpreted differently between these roles. While curse words can be a good indication of harassment, identification of a bully arguably requires more than these alone. Consider Table 1: both *B* and *A* use insults (lines 7–8), the message of *V* (line 6) might be considered as bullying in isolation, and having already determined *B*, the last sentence (line 10) can generally be regarded as a threat. In conclusion, the full scope of the task is complex; it could

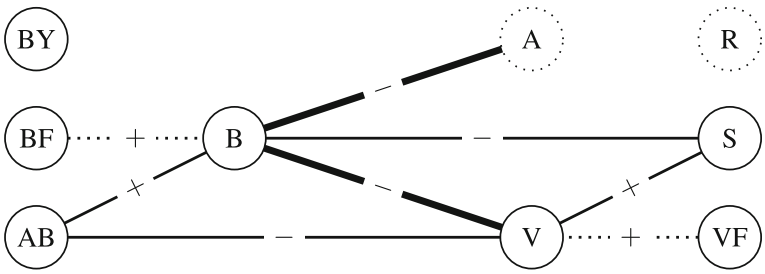

Fig. 1 Role graph of a bullying event. Each vertex represents an actor, labeled by their role in the event: bully (*B*), victim (*V*), bystander (*BY*), reinforcer (*AB*), assistant (*BF*), defender (*S*), reporter (*S*), accuser (*A*), and friend (*VF*). Each edge indicates a stream of communication, labeled by whether this is positive (+) or negative (−) in nature, and its strength indicating the frequency of interaction. Dotted edges indicate nonparticipation in the event, and vertices those added by Xu et al. (2012) to account for social-media-specific roles

Table 1 Fictional example of a cyberbullying conversation. Lines represent sequential turns

Line	Role	Message	Bully	Type
1	V	me and my friends hanging out tonight! :)		Neutral
2	B	@V lol b*tch, you dont have any friends.. ur fake as sh*t	✔	Curse, insult
3	AB	@B haha word, shes so sad	✔	Encouragement
4	VF	@V you know it girl		
5	S	@V dont listen to @B, were gonna have fun for sure!		Defense
6	V	@B shut up @B!! nobody asked your opinion!!!!		Defense
7	A	@B you are a f*cking bully, go outside or smt		Insult
8	B	@V @S haha you all so dumb, just kill yourself already!	✔	Insult, curse
9	A, R	@B shut up or ill report you		
10	B	@A u gonna cry? go ahead, see what happens tomorrow!	✔	Threat

Roles are noted as described on Page 4 (under the eponymous paragraph), if the message can be considered bullying by ✔, and types according to Van Hee et al. (2015)

have a temporal-sequential character, would benefit from determining actors and their interactions, and then should have some sense of severity as well (e.g. distinguish bullying from teasing).

1.6 Our contributions

Surprisingly, a significant amount of work on the task does not collect (or use) data that allows for the inference of such features (which we will further elaborate on in Sect. 3). To confirm this, we reproduce part of the previous cyberbullying detection research on different sources. Predictions made by current automatic methods for cyberbullying classification are demonstrated not to reflect the above-described task complexity; we show performance drops across different training domains, and give insights into content feature importance and limitations. Additionally, we report on reproducibility issues in state-of-art work when subjected to our evaluation. To facilitate future reproduction, we will provide all code open-source, including dataset readers, experimental code, and qualitative analyses.[3] Finally, we present a method to collect crowdsourced cyberbullying data in an experimental setting. It grants control over the size and richness of the data, does not invade privacy, nor rely on external parties to facilitate data access. Most importantly, we demonstrate that it successfully increases classifier performance. With this work, we provide suggestions on improving methodological rigor and hope to aid the community in a more realistic evaluation and implementation of this task of societal importance.

[3] Available at https://github.com/cmry/amica.

2 Related work

The task of detecting cyberbullying content can be roughly divided into three categories. First, research with a focus on *binary* classification, where it is only relevant if a message contains bullying or not. Second, more *fine-grained* approaches where the task is to determine either the role of actors in a bullying scenario or the content type (i.e., different categories of bullying). Both binary and fine-grained approaches predominantly focus on text-based features. Lastly, *meta-data* approaches that take more than just message content into account; these might include profile, network, or image information. Here, we will discuss efforts relevant to the task of cyberbullying classification within these three topics. We will predominantly focus on work conducted on openly available data, and those that report (positive) F_1-scores, to promote fair comparisons.[4] For an extensive literature review and a detailed comparison of different studies, see Rosa et al. (2019). Finally, a significant portion of our research pertains to generalizability, and therefore the field of domain adaptation. We will discuss its previous observations related text classification specifically, and their relevance to (future) research on cyberbullying detection.

2.1 Binary classification

One of the first traceable suggestions for applying text mining specifically to the task of cyberbullying detection is made by Kontostathis et al. (2010), who note that Yin et al. (2009) previously tried to classify online harassment on the CAW 2.0 dataset.[5] In the latter research, Yin et al. already state that the ratio of documents with harassing content to typical documents is challengingly small. Moreover, they foresee several other critical issues with regards to the task: a lack of positive instances will make detecting characteristic features a difficult task, and human labeling of such a dataset might have to face issues of ambiguity and sarcasm that are hard to assess when messages are taken out of conversation context. Even with very sparse datasets (with less than 1% positive class instances), the harassment classifier outperforms the random baseline using tf·idf, pronoun, curse word, and post similarity features.

Following up Yin et al. (2009), Reynolds et al. (2011) note that the CAW 2.0 dataset is generally unfit for cyberbullying classification: in addition to lacking bullying labels (it only provides harassment labels), the conversations are predominantly between adults. Their work, along with Bayzick et al. (2011), is a first effort to create datasets for cyberbullying classification through scraping the question-answering website Formspring.me, as well as Myspace.[6] In contrast with similar research, they aim to use textual features while deliberately avoiding Bag-

[4] Unfortunately, numerous (recent) work on cyberbullying detection seems not to report such F_1-scores (in favor of accuracy), is limited to criticized datasets with high baseline scores (such as the CAW datasets) or does not show enough methodological rigor—some are therefore not included in this overview.

[5] Data has been made available at http://caw2.barcelonamedia.org.

[6] Data has been made available at http://www.chatcoder.com/DataDownload.

of-Words (BoW) features. Through a curse word dictionary and custom severity annotations, they construct several metrics for features related to these "bad" words. In their more recent paper, Kontostathis and Reynolds (2013) redid analyses on the KON_FRM set, primarily focusing on the contribution curse words have in the classification of bullying messages. By forming queries from curse word dictionaries, they show that there is no one combination which retrieves all. However, by capturing them in an Essential Dimensions of Latent Semantic Indexing query vector averaged over known bullying content—classifying the top-*k* (by cosine similarity) as positive—they show a significant Average Precision improvement over their baseline.

More recent efforts include Bretschneider et al. (2014), who combined word normalization, Named Entity Recognition to detect person-specific references, and multiple curse word dictionaries (Noswearing.com 2016; Broadcasting Standards Authority 2013; Millwood-Hargreave 2000) in a rule-based pattern classifier, scoring well on Twitter data.[7] Our own work (Hee et al. 2015), where we collected a large dataset with posts from Ask.fm, used standard BoW features as a first test. Later, these were extended in Van Hee et al. (2018) with term lists, subjectivity lexicons, and topic model features. Recently popularized techniques of word embeddings and neural networks have been applied by Zhao et al. (2016); Zhao and Mao (2016) on XU_TREC, NAY_MSP and SUI_TWI, both resulting in the highest performance for those sets. Convolutional Neural Networks (CNNs) on phonetic features were applied by Zhang et al. (2016) and Rosa et al. (2018) investigate among others the same architecture on textual features in combination with Long Short Term Memory Networks (LSTMs). Both Rosa et al. (2018) and that of Agrawal and Awekar (2018) investigate the C-LSTM (Zhou et al. 2015), the latter includes Synthetic Minority Over-sampling Technique (SMOTE). However, as we will show in the current research, both of these works suffer from reproducibility issues. Finally, fuzzified vectors of top-*k* word lists for each class were used to conduct membership likelihood-based classification by Rosa et al. (2018) on KON_FRM, boosting recall over previously used methods.

2.2 Fine-grained classification

The common denominator of the previously discussed research was a focus on detecting single messages with evidence of cyberbullying per instance. As argued in Sect. 1.2, however, there are more textual cues to infer than merely if a message might be interpreted as bullying. The work of Xu et al. (2012) proposed to expand this binary approach with more fine-grained classification tasks by looking at *bullying traces*; i.e., the responses to a bullying incident. They distinguished two tasks based on keyword-retrieved (*bully*) Twitter data:[8] (1) a *role* labeling task, where semantic role labeling was then used to distinguish person-mention roles, and (2) the incorporation of sentiment to determine *teasing*, where despite high accuracy, 48% of the positive instances were misclassified.

[7] Data has been made available at http://www.ub-web.de/research.

[8] Data has been made available at http://research.cs.wisc.edu/bullying/data.html.

In our work, we extended this train of thought and demonstrated the difficulty of fine-grained classification of types of bullying (curse, defamation, defense, encouragement, insult, sexual, threat), and roles (harasser, bystander assistant, bystander defender, victim) with simple BoW and sentiment features—especially in detecting types (Hee et al. 2015; Van Hee et al. 2015). Later, this was further addressed in Van Hee et al. (2018) for both Dutch and English. Evaluated against a profanity (curse word lexicon) and word *n*-gram baseline, a multi-feature model (as discussed in Sect. 2.1) achieved the lowest error rates over (almost) all labels, for both bullying type and role classification. Lastly, Tomkins et al. (2018) also adapt fine-grained knowledge about bullying events in their socio-linguistic model; in addition to performing bullying classification, they find latent text categories and roles, partly relying on social interactions on Twitter. It thereby ties in with the next category of work: leveraging meta-data from the network the data is collected from.

2.3 Meta-data features

A notable, yet less popular aspect of this task is the utilization of a graph for visualizing potential bullies and their connections. This method was first adopted by Nahar et al. (2013), who use this information in combination with a classifier trained on LDA and weighted tf·idf features to detect bullies and victims on the CAW_* datasets. Work that more concretely implements techniques from graph theory is that of Squicciarini et al. (2015), who used a wide range of features: network features to measure popularity (e.g., degree centrality, closeness centrality), content-based features, (length, sentiment, offensive words, second-person pronouns), and incorporated age, gender, and number of comments. They achieved the highest performance on KON_FRM and BAY_MSP.

Work by Hosseinmardi et al. (2015) focuses on Instagram posts and incorporates platform-specific features retrieved from images and its network. They are the first to adhere to the literature more closely and define cyberagression (Kowalski et al. 2012) separately from cyberbullying, in that these are single negative posts rather than the repeated character of cyberbullying. They also show that certain LIWC (Linguistic Inquiry and Word Count) categories, such as death, appearance, religion, and sexuality, give a good indication of cyberbullying. While BoW features perform best, meta-data features (such as user properties and image content) in combination with textual features from the top 15 comments achieve a similar score. Cyberagression seems to be slightly easier to classify.

2.4 Domain adaptation

As the majority of the work discussed above focuses on a single corpus, a serious omission seems to be gauging how this influences model generalization. Many applications in natural language processing (NLP) are often inherently limited by expensive high-quality annotations, whereas unlabeled data is plentiful. Idiosyncrasies between source and target domains often prove detrimental to the performance of techniques relying on those annotations McClosky et al. (2006), Chan et al. (2006); Vilain et al. (2007) when applied in the wild. The field of

domain adaptation identifies tasks that suffer from such limitations, and aims to overcome them either in a supervised DauméIII (2009); Finkel et al. (2009) or unsupervised Blitzer et al. (2007); Jiang and Zhai (2007); Ma et al. (2014); Schnabel and Schütze (2014) way. For text classification, sentiment analysis is arguably closest to the task of cyberbullying classification Glorot et al. (2011); Chan et al. (2012); Pan et al. (2010). In particular as imbalanced data exacerbates generalization Li et al. (2012). However, while for sentiment analysis these issues are clearly identified and actively worked on, the same cannot be said for cyberbullying detection,[9] where concrete limitations have yet to be explored. We assume to find issues similar to those in sentiment analysis in the current task, as we will further discuss in the following section.

3 Task evaluation importance and hypotheses

The domain of cyberbullying detection is in its early stages, as can be seen in Table 2. Most datasets are quite small, and only a few have seen repeated experiments. Given the substantial societal importance of improving the methods developed so far, pinpointing shortcomings in the current state of research should assist in creating a robust framework under which to conduct future experiments—particularly concerning evaluating (domain) generalization of the classifiers. The latter of which, to our knowledge, none of the current research seems involved with. This is therefore the main focus of our work. In this section, we define three motivations for assessing this, and pose three respective hypotheses through which we will further investigate current limitations in cyberbullying detection.

3.1 Data scarcity

Considering the complexity of the social dynamics underlying the target of classification, and the costly collection and annotation of training data, the issue of data scarcity can mostly be explained with respect to the aforementioned restrictions on data access: while on a small number of platforms most data is accessible without any internal access (commonly as a result of optional user anonymity), it can be assumed that a significant part of actual bullying takes place 'behind closed doors'. To uncover this, one would require access to all known information within a social network (such as friends, connections, and private messages, including all meta-data). As this is unrealistic in practice, researchers rely on the small subset of publicly accessible data (predominantly text) streams. Consequently, most of the datasets used for cyberbullying detection are small and exhibit an extreme skew between positive and negative messages (as can be seen in Table 3). It is unlikely that these small sets accurately capture the language use on a given platform, and generalizable linguistic features of the bullying instances even less so. In line with

[9] One very recent exception to the latter can be found in Cheng et al. (2020). Their work introduces a novel domain adaptation technique, and demonstrates it to increase performance on two text classification tasks, one being cyberbullying detection.

Table 2 Overview of datasets for cyberbullying detection

Author	Other	Name	OS	Platform	Pos	Neg	Max	F_1
Yin et al. (2009)	Nahar et al. (2013)	CAW_KON	v	Kongregate	42	4802	Nahar et al. (2013)	.920
Yin et al. (2009)	Nahar et al. (2013)	CAW_SLS	v	Slashdot	60	4303	Nahar et al. (2013)	.920
Yin et al. (2009)	Nahar et al. (2013)	CAW_MSP	v	Myspace	65	1946	Nahar et al. (2013)	.920
Reynolds et al. (2011)	Kontostathis and Reynolds (2013), Squicciarini et al. (2015), Rosa et al. (2018, (2018)	KON_FRM	v	Formspring	369	3915	Rosa et al. (2018)	.848
Dinakar et al. (2011)	-	DIN_YTB	x	YouTube	2277	4500	Dinakar et al. (2011)	–
Bayzick et al. (2011)	Zhao and Mao (2016), Squicciarini et al. (2015)	BAY_MSP	v	Myspace	415	1647	Zhao and Mao (2016)	.776
Xu et al. (2012)	Zhao et al. (2016)	XU_TREC	v	Twitter	684	1762	Zhao et al. (2016)	.780
Dadvar (2014)	-	DDV_MSP	x	Myspace	311	8938	Dadvar (2014)	.350
Dadvar (2014)	-	DDV_YTB	x	YouTube	449	4177	Dadvar (2014)	.640
Bretschneider et al. (2014)	-	BRT_TWI	v	Twitter	220	5162	Bretschneider et al. (2014)	.726
Bretschneider et al. (2014)	-	BRT_TW2	v	Twitter	194	2599	Bretschneider et al. (2014)	.719
Hee et al. (2015)	Van Hee et al. (2018)	AMI_ASK	v	Ask.fm	3787	86419	Hee et al. (2015)	.642
Hosseinmardi et al. (2015)	Cheng et al. (2019)	HOS_INS	v	Instagram	567	1387	Cheng et al. (2019)	.783
Sui (2015)	Zhao and Mao (2016)	SUI_TWI	v	Twitter	2102	5219	Zhao and Mao (2016)	.719

Lists the authors of the initial sets (Author), if the set was used by other work (Other), a reference name (Name), if the data is publicly available (OS, including a link to the source), which platform it was extracted from (Platform), the number of reported cyberbullying instances (Pos) and of non-cyberbullying (Neg), a reference to the work achieving the highest score on the data (Max) and what the (positive) F_1-score was (F_1). Please note that the instance numbers are as reported in the original work, and may have deviated through time (such as Twitter sets, and Formspring)

Table 3 Corpus statistics for English and Dutch cyberbullying datasets, list number of positive (Pos, bullying) and negative (Neg, other) instances, Types (unique words), Tokens (total words), average number of tokens per message (Avg Tok/Msg), number of emojis and emoticons (Emote), and swear word occurrence per neutral (SweaN), and positive (SweaP) instance

	Pos	Neg	Types	Tokens	Avg Tok/Msg		Emote	SweaN	SweaP
D_{twB}	237	5258	12K	78K	14	($\sigma = 8$)	961	277	867
D_{frm}	1025	11,742	21K	348K	27	($\sigma = 29$)	3322	1228	2871
D_{msp}	426	1627	13K	803K	391	($\sigma = 285$)	931	1447	3730
D_{ytb}	417	3045	52K	827K	239	($\sigma = 252$)	3662	2606	8705
D_{ask}	5001	89,404	63K	1,017K	12	($\sigma = 23$)	17,362	4839	12,191
D_{twX}	281	4654	19K	86K	18	($\sigma = 8$)	1344	74	502
D_{tox}	15,279	144,226	220K	12,924K	81	($\sigma = 121$)	11,876	13,732	22,404
D_{ask_nl}	8675	70,557	58K	776K	10	($\sigma = 15$)	16,905	2025	2299
D_{sim_nl}	2330	2681	7K	55K	11	($\sigma = 16$)	434	682	194
D_{don_nl}	152	211	2K	7K	20	($\sigma = 24$)	33	47	19

Emojis were detected with https://github.com/NeelShah18/emot. Swears were detected with reference lists: for English these were taken from https://www.noswearing.com and the Dutch were manually composed.

domain adaptation research, we therefore anticipate that the samples are underpowered in terms of accurately representing the substantial language variation between platforms, both in normal language use and bullying-specific language use (Hypothesis 1).

3.2 Task definition

Furthermore, we argue that this scarcity introduces issues with adherence to the definition of the task of cyberbullying. The chances of capturing the underlying dynamics of *cyberbullying* (as defined in the literature) are slim with the message-level (i.e., using single documents only) approaches that the majority of work in the field has used up until now. The users in the collected sources have to be rash enough to bully in the open, and particular (curse) word usage that would explain the effectiveness of dictionary and BoW-based approaches in previous research. Hence, we also assume that the positive instances are biased; only reflecting a limited dimension of bullying (Hypothesis 2). A more realistic scenario—where characteristics such as repetitiveness and power imbalance are taken into consideration—would require looking at the interaction between persons, or even profile instances rather than single messages, which, as we argued, is not generally available. The work found in the meta-data category (Sect. 2.3) supports this argument with improved results using this information.

This theory regarding the definition (or operationalization) of this task is shared by Rosa et al., who pose that "*the most representative studies on automatic cyberbullying detection, published from 2011 onward, have conducted isolated*

online aggression classification" (Rosa et al. 2019, p. 341). We will mainly focus on the shared notion that this framing is limited to verbal aggression; however, our focus will empirically assess its overlap with data framed to solely contain online toxicity data (i.e., online / cyberagression) to find concrete evidence.

3.3 Domain influence

Enriching previous work with data such as network structure, interaction statistics, profile information, and time-based analyses might provide fruitful sources for classification and a correct operationalization of the task. However, they are also domain-specific, as not all social media have such a rich interaction structure. Moreover, it is arguably naive to assume that social networks such as Facebook (for which in an ideal case, all aforementioned information sources are available) will stay a dominant platform of communication. Recently, younger age groups have turned towards more direct forms of communication such as WhatsApp, Snapchat, or media-focused forms such as Instagram (Smith and Anderson 2018), and recently TikTok. This move implies more private and less affluent environments in which data can be accessed (resulting in even more scarcity), and that further development in the field requires a critical evaluation of the current use of the available features, and ways to improve cross-domain generalization overall. This work, therefore, does not disregard textual features; they would still need to be considered as the primary source of information, while paying particular attention to the issues mentioned here. We further try to contribute towards this goal and argue that crowdsourcing bullying content potentially decreases the influence of domain-specific language use, allows for richer representations, and alleviates data scarcity (Hypothesis 3).

4 Data

For the current research, we distinguish a large variety of datasets. For those provided through the AMiCA (Automatic Monitoring in Cyberspace Applications)[10] project, the *Ask.fm* corpus is partially available open-source,[11] and the *Crowdsourced* corpus will be made available upon request. All other sources are publicly available datasets gathered from previous research[12] as discussed in Sect. 2. Corpus statistics of all data discussed below can be found in Table 3. The sets' abbreviations, language (EN for English, NL for Dutch), and brief collection characteristics can be found below.

[10] http://www.amicaproject.be.

[11] https://osf.io/rgqw8/.

[12] These were collected as complete as possible. Twitter, in particular, has low recall; only an approximate of 60% of the tweets were retrieved. Such numbers are expected given the classification problem; people tend to remove harassing messages as was shown before by Xu et al. (2012).

4.1 AMiCA

Ask.fm (D_{ask}, D_{ask_nl}, EN, NL) were collected from the eponymous social network by Hee et al. (2015). Ask.fm is a question answering-style network where users interact by (frequently anonymously) asking questions on other profiles, and answering questions on theirs. As such, a third party cannot react to these question-answer pairs directly. The anonymity and restrictive interactions make for a high amount of potential cyberbullying. Profiles were retrieved through profile seed list, used as a starting point for traversing to other profiles and collecting all existing question-answer pairs for those profiles—these are predominantly Dutch and English. Each message was annotated with fine-grained labels (further details can be found in Van Hee et al. (2015)); however, for the current experiments these were binarized, with any form of bullying being labeled positive.

Donated (D_{don_nl}, NL) contains instances of (Dutch) cyberbullying from a mixture of platforms such as Skype, Facebook, and Ask.fm. The set is quite small; however, it contains several hate pages that are valuable collections of cyberbullying directed towards one person. The data was donated for use in the AMiCA project by previously bullied teens, thus forming a reliable source of gold standard, real-life data.

Crowdsourced (D_{sim_nl}, NL) originates from a crowdsourcing experiment conducted by Broeck et al. (2014), wherein 200 adolescents aged 14 to 18 partook in a role-playing experiment on an isolated SocialEngine[13] social network. Here, each respondent was given the account of a fictitious person and put in one of four roles in a group of six: a bully, a victim, two bystander-assistants, and two bystander-defenders. They were asked to read—and identify with—a character description and respond to an artificially generated initial post attributed to one of the group members. All were confronted with two initial posts containing either low- or high-perceived severity of cyberbullying.

4.2 Related work

Formspring (D_{frm}, EN) is taken from the research by Reynolds et al. (2011) and is composed of posts from Formspring.me, a question-answering platform similar to Ask.fm. As Formspring is mostly used by teenagers and young adults, and also provides the option to interact anonymously, it is notorious for hosting large amounts of bullying content (Binns 2013). The data was annotated through Mechanical Turk, providing a single label by majority vote for a question-answer pair. For our experiments, the question and answer pairs were merged into one document instance.

Myspace (D_{msp}, EN) was collected by Bayzick et al. (2011). As this was set up as an information retrieval task, the posts are labeled in batches of ten posts, and thus a single label applies to the entire batch (i.e., does it include cyberbullying). These were merged per batch as one instance and labeled accordingly. Due to this

[13] http://www.socialengine.com.

batching, the average tokens per instance are much higher than any of the other corpora.

Twitter (D_{twB}, EN) by Bretschneider et al. (2014) was collected from the stream between 20-10-2012 and 30-12-2012, and was labeled based on a majority vote between three annotators. Excluding re-tweets, the main dataset consists of 220 positive and 5162 negative examples, which adheres to the general expected occurrence rate of 4%. Their comparably-sized test set, consisting of 194 positive and 2699 negative examples, was collected by adding a filter to the stream for messages to contain any of the words `school`, `class`, `college`, and `campus`. These sets are merged for the current experiments.

Twitter II (D_{twX}, EN) from Xu et al. (2012) focussed on *bullying traces*, and was thus retrieved by keywords (`bully`, `bullying`), which if left unmasked generates a strong bias when utilized for classification purposes (both by word usage as well as being a mix of toxicity and victims). It does, however, allow for demonstrating the ability to detect bullying-associated topics, and (indirect) reports of bullying.

4.3 Experiment-specific

Ask.fm Context (C_{ask}, C_{ask}_nl, EN, NL)—the Ask.fm corpus was collected on profile level, but prior experiments have focused on single message instances (Van Hee et al. 2018). Here, we aggregate all messages for a single profile, which is then labeled as positive when as few as a single bullying instance occurs on the profile. This aggregation shifts the task of cyberbullying message detection to victim detection on profile level, allowing for more access to context and profile-level severity (such as repeated harassment), and makes for a more balanced set (1,763 positive and 6,245 negative instances).

Formspring Context (C_{frm}, EN)—similar to the Ask.fm corpus, was collected on profile level (Reynolds et al. 2011). However, the set only includes 49 profiles, some of which only include a single message. Grouping on full profile level would result in very few instances; thus, we opted for creating small 'context' in batches of five (of the same profile). Similar to the Ask.fm approach, if one of these messages contains bullying, it is labeled positive, balancing the dataset (565 positive and 756 negative instances).

Toxicity (D_{tox}, EN)—bashed on the Detox set from Wikimedia (Thain et al. 2017; Wulczyn et al. 2017), this set offers over 300k messages[14] of Wikipedia Talk comments with Crowdflower-annotated labels for toxicity (including subtypes).[15] Noteworthy is how *disjoint* both the task and the platform are from the rest of the corpora used in this research. While toxicity shares many properties with bullying, the focus here is on single instances of insults directed to anonymous people, who are most likely unknown to the harasser. Given Wikipedia as a source, the article and moderation focussed comments make it topically quite different from what one would expect on social media—the fundamental overlap being curse words, which

[14] From: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[15] See https://meta.wikimedia.org/wiki/Research:Detox/Data_Release for more information regarding operationalization of this dataset.

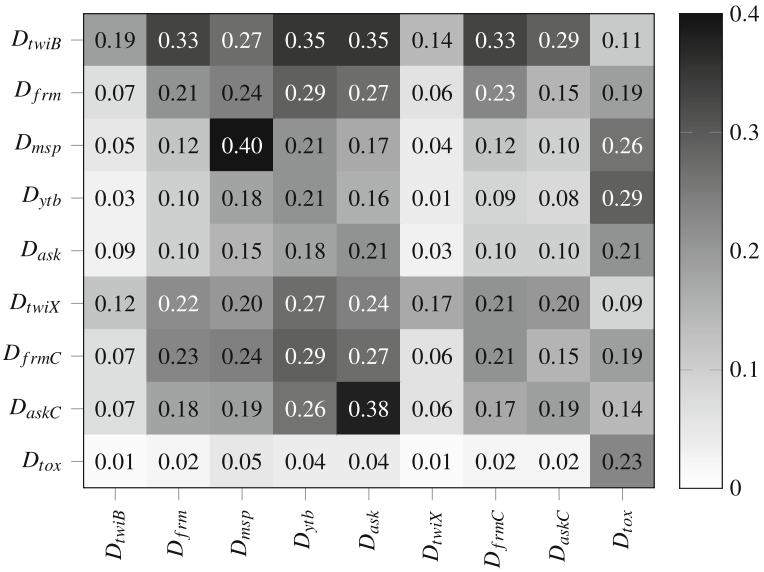

Fig. 2 Jaccard similarity between training sets (*y*-axis) and test sets (*x*-axis)

is only one of many dimensions to be captured to detect cyberbullying (as opposed to toxicity).

4.4 Preprocessing

All texts were tokenized using spaCy (Honnibal and Montani 2017).[16] No preprocessing was conducted for the corpus statistics in Table 3. All models (Sect. 5) applied lowercasing and special character removal only; alternative preprocessing steps proved to decreased performance (see Table 6).

4.5 Descriptive analysis

Both Table 3 and Fig. 2 illustrate stark differences; not only across domains but more importantly, between in-domain training and test sets. Most do not exceed a Jaccard similarity coefficient over 0.20 (Fig. 2), implying a large part of their vocabularies do not overlap. This contrast is not necessarily problematic for classification; however, it does hamper learning a general representation for the negative class. It also clearly illustrates how even more disjoint D_{twX} (collected by trace queries) and D_{tox} are from the rest of the corpora and splits. Finally, the descriptives (Table 3) further show significant differences in size, message length, class balance, and type/token ratios (i.e., writing level). In conclusion, it can be assumed that the language use in both positive as negative instances will vary

[16] https://spacy.io (`v2.0.5`).

significantly, and that it will be challenging to model in-domain, and generalize out-of-domain.

5 Experimental setup

We attempt to address the following hypotheses posited in Sect. 3:

- Hypothesis 1: As researchers can only rely on scarce data of public bullying, their samples are assumed to be underpowered in terms of accurately representing the substantial language variation between platforms, both in normal language use and bullying-specific language use.
- Hypothesis 2: Given knowledge from the literature, it is unlikely that single messages capture the full complexity of bullying events. Cyberbullying instances in the considered corpora are therefore expected to be largely biased, only reflecting a limited dimension of bullying.
- Hypothesis 3: With control over data generation and structure, crowdsourcing bullying content potentially decreases the influence of domain-specific language use, allows for richer representations, and alleviates data scarcity.

Accordingly, we propose five main experiments. Experiments I (Hypothesis 1) and III (Hypothesis 2) deal with the problem of generalizability, whereas Experiment II (Hypothesis 1) and V (Hypothesis 3) will both propose a solution for restricted data collection. Experiment IV will reproduce a selection of state-of-the-art models for cyberbullying detection and subject them to our cross-domain evaluation, to be compared against our baselines.

5.1 Experiment I: cross-domain evaluation

In this experiment, we introduce the cross-domain evaluation framework, which will be extended in all other experiments. For this, we initially perform a many-to-many evaluation of a given model (baseline or otherwise) trained individually on all available data sources, split in train and test. In later experiments, we extend this with a one-to-many evaluation. This setup implies that (i) we fit our model on some given corpus' training portion and evaluate prediction performance on all available corpora their test portions (many-to-many) individually. Furthermore, we (ii) fit on all corpora their train portions combined, and evaluate on all their test portions individually (one-to-many). In sum, we report on 'small' models trained on each corpus individually, as well as a 'large' one trained on them combined, for each test set individually.

For every experiment, hyper-parameter tuning was conducted through an exhaustive grid search, using nested cross-validation (with ten inner and three outer folds) on the training set to find the optimal combination of the given parameters. Any model selection steps were based on the evaluation of the outer folds. The best performing model was then refitted on the full training set (90% of the data) and applied to the test set (10%). All splits (also during cross-validation)

were made in a stratified fashion, keeping the label distributions across splits similar to the whole set.[17] Henceforth, all experiments in this section can be assumed to follow this setup.

The many-to-many evaluation framework intends to test Hypothesis 1 (Sect. 3.1), relating to language variation and cross-domain performance of cyberbullying detection. To facilitate this, we employ an initial `baseline` model: Scikit-learn's (Pedregosa et al. 2011) Linear Support Vector Machine (SVM) (Cortes and Vapnik 1995; Fan et al. 2008) implementation trained on binary BoW features, tuned using the grid shown in Table 3, based on Van Hee et al. (2018). Given its use in previous research, it should form a strong candidate against which to compare. To ascertain out-of-domain performance compared to this baseline, we report test score averages across all test splits, excluding the set the model was trained on (in-domain).

Consequently, we add an evaluation criterion to that of related work: when introducing a novel approach to cyberbullying detection, it should not only perform best in-domain for the majority of available corpora, but should also achieve the highest out-of-domain performance on average to classify as a robust method. It should be noted that the selected corpora for this work are not all optimally representative for the task. The tests in our experiments should, therefore, be seen as an initial proposal to improve the task evaluation.

5.2 Experiment II: gauging domain influence

In an attempt to overcome domain restrictions on language use, and to further solidify our tests regarding Hypothesis 1, we aim to improve the performance of our baseline models through changing our representations in three distinct ways: (i) merging all available training sets (as to simulate a large, diverse corpus), (ii) by aggregating instances on user-level, and (iii) using state-of-the-art language representations over simple BoW features in all settings. We define these experiments as such:

Volume and Variety Some corpora used for training are relatively small, and can thus be assumed insufficient to represent held-out data (such as the test sets). One could argue that this can be partially mitigated through simply collecting more data or training on multiple domains. To simulate such a scenario, we merge all available cyberbullying-related training splits (creating D_{all}), which then corresponds to the one-to-many setting of the evaluation framework. The hope is that corpora similar in size or content (the Twitter sets, Ask.fm and Formspring, YouTube and Myspace) would benefit from having more (related) data available. Additionally, training a large model on its entirety facilitates a catch-all setting for assessing the average cross-domain performance of the full task (i.e. across all test sets when trained on all available corpora). This particular evaluation will be used in Experiment IV (replication) for model comparison.

Context change Practically all corpora, save for MySpace and YouTube, have annotations based on short sentences, which is particularly noticeable in Table 3. This one-shot (i.e., based on a single message) method of classifying cyberbullying

[17] Indices (similar to any other random components) were fixed by the same seed for all experiments.

provides minimal content (and context) to work with. It does therefore not follow the definition of cyberbullying—as previously discussed in Sect. 3.2. As a preliminary simulation[18] of adding (richer) context, we merge the profiles of D_{ask} and (batches of) D_{frm} into single context instances (creating C_{ask} and C_{frm}, see Sect. 4). This allows us to compare models trained on larger contexts directly to that of single messages, and evaluate how context restrictions affect performance on the task in general, as well as cross-domain.

Improving representations Pre-trained word embeddings as language representation have been demonstrated to yield significant performance gains for a multitude of NLP-related tasks (Collobert et al. 2011). Given the general lack of training data—including negative instances for many corpora—word features (and weightings) trained on the available data tend to be a poor reflection of the language use on the platform itself, let alone other social media platforms. Therefore, pre-trained semantic representations provide features that in theory, should perform better in cross-domain settings. We consider two off-the-shelf embedding models per language that are suitable for the task at hand: for English, averaged 200-dimensional GloVe (Pennington et al. 2014) vectors trained on Twitter,[19] and DistilBERT (Sanh et al. 2019) sentence embeddings (Devlin et al. 2018).[20] For Dutch, `fastText` embeddings (Bojanowski et al. 2017) trained on Wikipedia[21] and `word2vec` (Mikolov et al. 2013) embeddings[22] (Tulkens et al. 2016) trained on the COrpora from the Web (COW) corpus (Schäfer and Bildhauer 2012) embeddings. The GLoVe, `fastText`, and `word2vec` embeddings were processed using Gensim[23] (Řehůřek et al. 2010).

As an additional baseline for this section, we include the Naive Bayes Support Vector Machine (NBSVM) from Wang et al. (2012), which should offer competitive performance on text classification tasks.[24] This model also served as a baseline for the Kaggle challenge related to D_{tox}.[25] NBSVM uses tf·idf-weighted uni and bigram features as input, with a minimum document frequency of 3, and corpus prevalence of 90%. The idf values are smoothed and tf scaled sublinearly $(1 + \log(\text{tf}))$. These are then weighted by their log-count ratios derived from Multinomial Naive Bayes.

Tuning of both embeddings and NB representation classifiers is done using the same grid as Table 4, however replacing C with [1, 2, 3, 4, 5, 10, 25, 50, 100, 200, 500]. Lastly, we opted for Logistic Regression (LR), primarily as this was used in the NBSVM implementation mentioned above, as well as `fastText`. Moreover, we found SVM using our grid to perform marginally worse using these features, and

[18] Preferably, one would want to collect data on profile level by design. The corpora available were not specifically collected this way, making our set-up an approximation of such a setting.

[19] https://nlp.stanford.edu/projects/glove/ (`v1.2`).

[20] https://github.com/huggingface/transformers (`1d646ba`).

[21] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md (`2665eac`).

[22] https://github.com/clips/dutchembeddings (`1e3d528`).

[23] https://radimrehurek.com/gensim/index.html (`v3.4`).

[24] The implementations for these models can be found in our repository.

[25] https://kaggle.com/jhoward/nb-svm-strong-linear-baseline/notebook.

Table 4 SVM baseline and NBSVM grid used in hyper-parameter search

Part	Params	Values
BoW	range level	(1, 1), (1, 2), (1, 3) words
SVM	weight y loss C	default, balanced hinge, square hinge 1e−, 1e−2, …, 1e2, 1e3

fine-tuning DistilBERT using a fully connected layer (similar to e.g. Sun et al. (2019)) to yield similar performance. The embeddings were not fine-tuned for the task. While this could potentially increase performance, it complicates direct comparison to our baselines—we leave this for Experiment IV.

5.3 Experiment III: aggression overlap

In previous research using fine-grained labels for cyberbullying classification (e.g., Van Hee et al. (2018)) it was observed that cyberbullying classifiers achieve the lowest error rates on blatant cases of aggression (cursing, sexual talk, and threats), an idea that was further adopted by Rosa et al. (2019). To empirically test Hypothesis 2 (see Sect. 3.2)—related to the bias present in the available positive instances—we adapt the idea of running a profanity baseline from this previous work. However, rather than relying on look-up lists containing profane words, we expand this idea by training a separate classifier on toxicity detection (D_{tox}) and seeing how well this performs on our bullying corpora (and vice-versa). For the corpora with fine-grained labels, we can further inspect and compare the bullying classes captured by this model.

We argue that high test set performance overlap of a toxicity detection model with models trained on cyberbullying detection gives strong evidence of nuanced aspects of cyberbullying not being captured by such models. Notably, in line with Rosa et al. (2019), that the current operationalization does not significantly differ from the detection of online aggression (or toxicity)—and therefore does not capture actual cyberbullying. Given enough evidence, both issues should be considered as crucial points of improvement for the further development of classifiers in this domain.

5.4 Experiment IV: replicating state-of-the-art

For this experiment, we include two architectures that achieved state-of-the-art results on cyberbullying detection. As a reference neural network model for language-based tasks, we used a Bidirectional (Schuster and Paliwal 1997; Baldi et al. 1999) Long Short-Term Memory network (Hochreiter and Schmidhuber 1997; Gers et al. 2002) (BiLSTM), partly reproducing the architecture from Agrawal and Awekar (2018). We then attempt to reproduce the Convolutional Neural Network (CNN) (Kim 2014) used in both Rosa et al. 2018) and Agrawal and Awekar (2018), and the Convolutional LSTM (C-LSTM) (Zhou et al. (2015) used in Rosa et al. (2018). As Rosa et al. (2018) do not report essential implementation details for

these models (batch size, learning rate, number of epochs), there is no reliable way to reproduce their work. We will, therefore, take Agrawal and Awekar (2018) their implementation for the BiLSTM and CNN as the initial setup. Given that this work is available open-source, we run the exact architecture (including SMOTE) in our Experiment I and II evaluations. The architecture-specific details are as follows:

Reproduction We initially adopt the basic implementation[26] by Agrawal and Awekar (2018): randomly initialized embeddings with a dimension of 50 (as the paper did not find significant effects of changing the dimension, nor initialization), run for 10 epochs with a batch size of 128, dropout probability of 0.25, and a learning rate of 0.01. Further architecture details can be found in our repository.[27] We also run a variant with SMOTE on, and one from the provided notebooks directly.[28] This and following neural models were run on an NVIDIA Titan X Pascal, using Keras (Chollet et al. 2015) with Tensorflow (Abadi et al. 2015) as backend.

BiLSTM For our own version of the BiLSTM, we minimally changed the architecture from Agrawal and Awekar (2018), only tuning using a grid on batch size [32, 64, 128, 256], embedding size [50, 100, 200, 300], and learning rate [0.1, 0.01, 0.05, 0.001, 0.005]. Rather than running for ten epochs, we use a validation split (10% of the train set) and initiate early stopping when the validation loss does not go down after three epochs. Hence—and in contrast to earlier experiments—we do not run the neural models in 10-fold cross-validation, but a straightforward 2-fold train and test split where the latter is 10% of a given corpus. Again, we are predominantly interested in confirming statements made in earlier work; namely, that for this particular setting tuning of the parameters does not meaningfully affect performance.

CNN We use the same experimental setup as for the BiLSTM. The implementations of Agrawal and Awekar (2018); Rosa et al. (2018) use filter window sizes of 3, 4, and 5—max pooled at the end. Given that the same grid is used, the word embedding sizes are varied and weights trained (whereas Rosa et al. (2018) use 300-dimensional pre-trained embeddings). Therefore, for direct performance comparisons, Agrawal and Awekar (2018) their results will be used as a reference. As CNN-based architectures for text classification are often also trained on character level, we include a model variant with this input as well.

C-LSTM For this architecture, we take an open-source text classification survey implementation.[29] This uses filter windows of [10, 20, 30, 40, 50], 64-dimensional LSTM cells and a final 128 dimensional dense layer. Please refer to our repository for additional implementation details—for this and previous architectures.

[26] https://github.com/sweta20/Detecting-Cyberbullying-Across-SMPs/blob/master/DNNs.ipynb.

[27] https://github.com/cmry/amica/blob/master/neural.py.

[28] Note this is for testing reproduction only, as it is not subjected to the same evaluation framework.

[29] https://github.com/bicepjai/Deep-Survey-Text-Classification/.

5.5 Experiment V: crowdsourced data

Following up on the proposed shortcomings of the currently available corpora in Hypotheses 1 and 2, we propose the use of a crowdsourcing approach to data collection. In this experiment, we will repeat Experiment I and II with the best out-of-domain classifier from the above evaluations with three (Dutch[30]) datasets: D_{ask_nl}; the Dutch part of the Ask.fm dataset used before, D_{sim_nl}; our synthetic, crowdsourced cyberbullying data, and lastly D_{don_nl}; a small donated cyberbullying test set with messages from various platforms (full overview and description of these three can be found in Sect. 4). The only notable difference to our setup for this experiment is that we never use D_{don_nl} as training data. Therefore rather than D_{all}, the Ask.fm corpus is merged with the crowdsourced cyberbullying data to make up the D_{comb} set.

6 Results and discussion

We will now cover results per experiment, and to what extent these provide support for the hypotheses posed in Sect. 3. As most of these required backward evaluation (e.g., Experiment III was tested on sets from Experiment I), the results of Experiment I-III are compressed in Table 5. Table 7 comprises the *Improving Representations* part of Experiment II (under 'word2vec' and 'DistilBERT') along with the preprocessing results effect of our baselines. The results of Experiment V can be found in Table 8. For brevity of reporting, the latter two only report on the in-domain scores, and feature the out-of-domain *averages* for the D_{all} models for comparison, and D_{tox} averages in Table 7.

6.1 Experiment I

Looking at Table 5, the upper group of rows under T1 represents the results for Experiment I. We posed in Hypothesis 1 that samples are underpowered regarding their representation of the language variation between platforms, both for bullying and normal language use. The data analysis in Sect. 4.5 showed minimal overlap between domains in vocabulary and notable variances in numerous aspects of the available corpora. Consequently, we raised doubts regarding the ability of models trained on these individual corpora to generalize to other corpora (i.e., domains).

Firstly, we consider how well our `baseline` performed on the *in-domain test sets*. For half of the corpora, it achieves the highest performance on these specific sets (i.e., the test set portion of the data the model was trained on). More importantly, this entails that for four of the other sets, models trained on other corpora perform equal or better. Particularly the effectiveness of D_{ask} was in some cases surprising; the YouTube corpus by Dadvar et al. (2014) (D_{ytb}), for example, contains much longer instances (see Table 2).

[30] On account of the synthetic data being available in Dutch only. Experiment III was not repeated as there is no equivalent toxicity dataset available in this language.

Table 5 Cross-corpora positive class F_1 scores for Experiment I (T1), II (T2), and III (T3)

Train	T1						Avg	T2		T3
	D_{twB}	D_{frm}	D_{msp}	D_{ytb}	D_{ask}	D_{twX}		C_{frm}	C_{ask}	D_{tox}
D_{twB}	.417	.308	.000	.122	.298	.051	.153	.131	.158	.349
D_{frm}	.423	.454	.042	.379	.418	.041	.321	.682	.259	.465
D_{msp}	.120	.176	**.941**	.324	.168	.043	.197	.364	.185	.185
D_{ytb}	.074	.160	.375	.365	.138	.000	.183	.338	.197	.140
D_{ask}	.493	.444	.211	**.421**	**.561**	.139	.351	.389	.357	.584
D_{twX}	.049	.131	.184	.175	.077	**.508**	.205	.496	.325	.082
D_{all}	**.524**	**.473**	**.941**	.397	.553	.194	**.557**	**.780**	.570	.587
C_{frm}	.152	.253	.143	.286	.136	.126	.214	.758	.400	.372
C_{ask}	.286	.237	.359	.244	.356	.107	.310	.582	**.579**	.280
D_{tox}	.343	.373	.449	.335	.443	.149	.389	.628	.539	**.806**

Models are fitted on the training proportion of the corpora row-wise, and tested column-wise. The out-of-domain average (Avg) excludes test performance of the parent training corpus. The best overall test score is noted in bold, the best out-of-domain performance in gray

It must be noted though, that the baseline was selected from work on the Ask.fm corpus (Van Hee et al. 2018). This data is also one of the more diverse datasets (and largest) with exclusively short messages; therefore, one could assume a model trained on this data would work well on both longer and shorter instances. It is however also likely that particularly this baseline (binary word features) trained on this data therefore enforces the importance of more shallow features. This will be further explored in Experiments II and III.

For Experiment I, however, our goal was to assess the out-of-domain performance of these classifiers, not to maximize performance. For this, we turn to the *Avg* column in Table 5. Between the *top portion* of the Table, the D_{ask} model performs best across all domains (achieving highest on three, as mentioned above). The second-best model is trained on the Formspring data from Reynolds et al. (2011) (D_{frm}), akin to Ask.fm as a domain (question-answer style, option to post anonymously). It can be observed that almost all models perform worst on the 'bullying traces' Twitter corpus by Xu et al. (2012), which was collected using queries. This result is relatively unsurprising, given the small vocabulary overlaps with its test set shown in Fig. 2. We also confirm in line with Reynolds et al. (2011) that the CAW data from Bayzick et al. (2011) is unfit as a bullying corpus; achieving significant positive F_1-scores with a baseline, generalizing poorly and proving difficult as a test set.

Additionally, we observe that even the best performing models yield between .1 and .2 lower F_1 scores on other domains, or a $15-30\%$ drop from the original score. To explain this, we look at how well important features generalize across test sets. As our `baseline` is a Linear SVM, we can directly extract all grams with positive coefficients (i.e., related to bullying). Figure 3 (right) shows the frequency of the top 5000 features with the highest coefficient values. These can be observed to follow a Zipfian-like distribution, where the important features most frequently

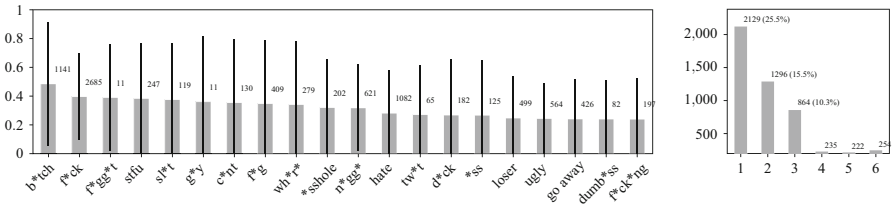

Fig. 3 Left: Top 20 *test set* words with the highest average coefficient values across all classifiers (minus the model trained on D_{tox}). Error bars represent standard deviation. Each coefficient value is only counted once per test set. The frequency of the words is listed in the annotation. Right: Test set occurrence frequencies (and percentages) of the top 5000 highest absolute feature coefficient values

occur in one test set (25.5%) only, which quickly drops off with increasing frequency. Conversely, this implies that over 75% of the top 5,000 features seen during training do not occur in any test instance, and only 3% generalize across all sets. This coverage decreases to roughly 60% and 4% respectively for the top 10,000, providing further evidence of the strong variation in predominantly bullying-specific language use.

Figure 3 (left) also indicates that the coefficient values are highly unstable across test sets, with most having roughly a 0.4 standard deviation. Note that these coefficient values can also flip to negative for particular sets, so for some of the features, the range goes from associated with the other class to highly associated with bullying. Given the results of Table 5 and Fig. 3, we can conclude that our `baseline` model shows not to generalize out-of-domain. Given the quantitative and qualitative results reported on in this Experiment, this particular setting partly supports Hypothesis 1.

6.2 Experiment II

The results for this experiment can be predominantly found in Table 5 (middle and lower parts, and T2 in particular), and partly in Table 7 (word2vec, DistilBERT). In this experiment, we seek to further test Hypothesis 1 by employing three methods: merging all cyberbullying data to increase *volume and variety*, aggregating on context level for a *context change*, and *improving representations* through pre-trained word embedding features. These are all reasonably straightforward methods that can be employed in an attempt to mitigate data scarcity.

Volume and variety The results for this part are listed under D_{all} in Table 5. For all of the following experiments, we now focus on the full results table (including that of Experiment I) and see which individual classifiers generalize best across all test sets (highlighted in gray). The *Avg* column shows that our 'big' model trained on all available corpora[31] achieves second-best performance on half of the test sets and best on the other half. More importantly, it has the highest average out-of-

[31] This average excludes toxicity data from D_{tox}, which we found when added to substantially decrease performance on all domains, except for D_{twX} and C_{frm}. Note that it also includes scopes from the *context change* experiment.

domain performance, without competition on any test set. These observations imply that for the `baseline` setting, an ensemble model of different smaller classifiers should not be preferred over the big model. Consequently, it can be concluded that collecting more data does seem to aid the task as a whole.

However, a qualitative analysis of the predictions made by this model clearly shows lingering limitations (see Table 6). These three randomly-picked examples give a clear indication of the focus on blatant profanity (such as `d*ck`, `p*ss`, and `f*ck`). Especially combinations of words that in isolation might be associated with bullying content (`leave`, `touch`) tend to confuse the model. It also fails to capture more subtle threats (`skull drag`) and infrequent variations (`h*`). Both of these structural mistakes could be mitigated by providing more context that potentially includes either more toxicity or more examples of neutral content to decrease the impact of single curse words—hence, the next experiment.

Context change As for access to context scopes, we are restricted to the Ask.fm and Formspring data (C_{frm} and C_{ask} in Table 5). Nevertheless, in both cases, we see a noticeable increase for in-domain performance: a positive F_1 score of .579 for context scope versus .561 on Ask.fm, and .758 versus .454 on Formspring respectively. This increase implies that considering message-level detection for both individual sets should be preferred. On the other hand, however, these longer contexts do perform worse on out-of-domain sets. This can be partly explained due the fact that including more data (therefore moving the data to profile, or conversation level) shifts the task to identifying bullying conversations, or profiles of victims. While variation will be higher, chances also increase that multiple single bullying messages will be captured in a one context. This would therefore allow to learn the distinction between a profile or conversation with predominately neutral messages including a single toxic message—which might therefore be harmless, to one where there are multiple toxic messages, increasing the severity.

The change in scope clearly influences which features are deemed important. On manual inspection, averaging feature importances of all `baseline` models on their in-domain test sets, the top 500 most important features consist for 63% of profane words. For the models trained on Ask.fm and Formspring specifically (D_{ask} and D_{frm}), this is an average of 42%. Strikingly, for the models trained on context scopes (C_{frm} and C_{ask}), this percentage significantly reduced to 11%; many of their important bi-gram features include `you`, topics such as `dating`, `boys`, `girls`, and `girlfriend` occur, yet also positive words such as (`are`) `beautiful`—the latter of which could indicate messages from friends (defenders). This change is to an extent expected as by changing the scope, the task shifts to classifying profiles that are bullied, thus showing more diverse bullying characteristics.

These results provide evidence for extending classification to contexts to be a worthwhile platform-specific setting to pursue. However, we can conversely draw the same conclusions as Experiment I; that including direct context does not overcome the tasks general domain limitations, therefore further supporting Hypothesis 1. A plausible solution to this could be improving upon the BoW features by relying on more general representations of language, as found in word embeddings.

Table 6 Examples of uni-gram weights according to the `baseline` SVM trained D_{all}, tested on D_{twB} and D_{ask}

y	$\hat{y}$	D_{twB}	D_{ask}
👍	👎	about to leave this school library	bigerrr ? how much ? its gon na
		and take my *ss homeeee	touch the sky ? a wonder d*ck ?
👎	👎	you p*ss me off so much .	r u a r*t*rd liam mate f*ck off
👎	👍	@username i will skull drag you	h* of me xoxoxoxoxoxoox
		across campus .	

Words in red are associated with bullying, words in green with neutral content. The color intensity is derived from the strength of the SVM coefficients per feature (most are near zero). Black boxes indicate OOV words. Labels are divided between the gold standard (y) and predicted ($\hat{y}$) labels, 👎 for bullying content, 👍 for neutral

Improving representations The aim of this experiment was to find (out-of-the-box) representations that would improve upon the simple BoW features used in our `base-line` model (i.e., achieving good in-domain performance as well as out-of-domain generalization). Table 7 lists both of our considered baselines, tested under different preprocessing methods. These are subsequently compared against the two different embedding representations.

For preprocessing, several levels were used: the default for all models being (1) lowercasing only, then either (2) removal of special characters, or (3) lemmatization and more appropriate handling of special characters (e.g., splitting `#word` to prepend a hashtag token) were added. The corresponding results in Table 7 do not reveal an unequivocal preprocessing method for either the BoW `baseline` or NBSVM. While the latter achieves highest out-of-domain generalization with thorough preprocessing ('+preproc', .566 positive F_1), the `baseline` model achieves best in-domain performance on five out of nine corpora, and an on-par out-of-domain average (.566 versus .561) with simple cleaning ('+clean').

According to our criterion proposed in Sect. 5.1, a method that performs well both in- and out-of-domain should be preferred. The current consideration of preprocessing methods illustrates how this stricter evaluation criterion used in this experiment potentially yields different overall results in contrast to evaluating in-domain only, or focusing on single corpora. Conversely, we opted for simple cleaning throughout the rest of our experiment (as mentioned in Sect. 4.4), given its consistent performance for both baselines.

The embeddings chosen for this experiment do not seem to provide representations that yield an overall improvement for the classification performance of our Logistic Regression model. Surprisingly, however, DistilBERT does yield significant gains over our baseline for the conversation-level corpus of Ask.fm (.629 positive F_1 over .579). This might imply that such representations would work well on more (balanced) data. While we did not see a significant effect on performance with shallowly fine-tuning DistilBERT, more elaborate fine-tuning would be a required point of further investigation before drawing strong conclusions.

Table 7 Overview of different feature representations (Repr) for Experiment I and II

Repr	T1						Avg	T2		T3
	D_{twB}	D_{frm}	D_{msp}	D_{ytb}	D_{ask}	D_{twX}		C_{frm}	C_{ask}	D_{tox}
`baseline`	**.417**	.454	**.941**	.365	.561	.508	.557	.758	.579	.806
+ clean	.408	**.477**	.927	.354	**.562**	**.517**	.561	**.764**	.592	**.807**
+ preproc	.345	.426	.929	**.377**	.506	.293	.512	.600	.582	.734
NBSVM	.364	.462	.929	.231	.508	.469	.542	.635	.592	.779
+ clean	.410	.456	.940	.211	.541	.467	.563	.641	.596	.747
+ preproc	.318	.466	.907	.320	.480	.305	**.566**	.532	.597	.756
word2vec	.368	.394	.860	.338	.304	.323	.366	.698	.572	.634
DistilBERT	.377	.336	.697	.296	.369	.435	.402	.598	**.629**	.642

The '+' parts show performance for preprocessing: removing all special characters (clean), and more sophisticated handling of social media tags and emojis (preproc). Their in-domain positive class F_1 scores for Experiment I (T1) and II (T2), and the out-of-domain average (Avg) for D_{all}. Baseline scores are from Table 5

Moreover, given that we restricted our embeddings to averaged representations on document-level for word2vec, and the sentence representation token for BERT (following common practices), numerous settings remain unexplored. While both (i.e., fine-tuning and alternate input representations) of such potential improvements would certainly merit further exploration in future work focused on optimization, this is out of scope of the current research. Similarly, embeddings trained on a similar domain would be more ideal to represent our noisy data; we settled for strictly off-the-shelf ones that at minimum included web content, and a large vocabulary.

Therefore, we conclude that no alternative (out-of-the-box) baselines seem to clearly outperform our BoW baseline. We previously eluded to the effectiveness of binary BoW representations in previous work, and argued this being a result of capturing blatant profanity. We will further test this in the next experiment.

6.3 Experiment III

Here, we investigate Hypothesis 2: the notion that positive instances across all cyberbullying corpora are biased, and only reflect a limited dimension of bullying. We have already found strong evidence for this in the previous Experiments I and II, Fig. 3, Table 6, and manual analyses of top features all indicated toxicity to be consistent top-ranking features. To add more empirical evidence to this, we trained models on toxicity, or cyber aggression, and tested them on bullying data (and vice-versa)—providing results on the overlap between the tasks. The results for this experiment can be found in the lower end of Table 5, under D_{tox} and T3.

It can be noted that there is a substantial gap in performance between the cyberbullying classifiers (using D_{all} as reference) performance on the D_{tox} test set

and that of the toxicity model (positive F_1 score of .587 and .806 respectively). More strikingly, however, the other way around, toxicity classifiers perform second-best on the out-of-domain averages (Avg in Table 5). In the context scopes (C_{frm} and C_{ask}) it is notably close, and for other sets relatively close, to the in-domain performance.

Cyberbullying detection should include detection of toxic content, yet also perform on more complex social phenomena, likely not found in the Wikipedia comments of the toxicity corpus. It is therefore particularly surprising that it achieves higher out-of-domain performance on cyberbullying classification than all individual models using BoW features to capture bullying content. Only when all corpora are combined, the D_{all} classifier performs better than the toxicity model. This observation combined with previous results provides significant evidence that a large part of the available cyberbullying content is not complex, and current models to only generalize to a limited extent using predominately simple aggressive features, supporting Hypothesis 3.

6.4 Experiment IV

So far, we have attempted to improve a straight-forward baseline that was trained on binary features with several different approaches. While changes in data (representations) seem to have a noticeable effect on performance (increasing the amount of messages per instance, merging all corpora), none of the experiments with different feature representations have had an impact. With the current experiment, we had hoped to leverage earlier state-of-the-art architectures by reproducing their methodology and subjecting our evaluation framework.

As can be inferred from Table 8, our baselines outperform these neural techniques on almost all in-domain tests, as well as the out-of-domain averages. Having strictly upheld the experimental set-up from Agrawal and Awekar (2018) and as close as possible that of Rosa et al. (2018), we can conclude that—under stricter evaluation—there is sufficient evidence that these models do not provide state-of-the-art results on the task of cyberbullying.[32] Tuning these networks (at least in our set-up) does not seem to improve performance, rather decrease it. This indicates that the validation set on which early stopping is conducted is often not representative to the test set. Parameter tuning on this set is consequently sensitive to overfitting; an arguably unsurprising result given the size of the corpora.

Some further noteworthy observations can be made related to the performance of the CNN architecture, achieving quite significant leaps on word level (for D_{twB}) and character level (for C_{ask}). Particularly the conversation scopes (C, with a

[32] Upon acquiring the results of the replication of Agrawal and Awekar (2018) (in particular failing to replicate the effect of the paper's oversampling) we investigated the provided code and notebooks. It is our understanding that oversampling before splitting the dataset into training and test sets causes the increase in performance; we measured overlap of positive instances in these splits and found no unique test instances. Furthermore, after re-running the experiments directly from the notebooks with the oversampling conducted post-split, the effect was significantly decreased (similar to our results in Table 7). The authors were contacted with our observations in March 2019, and have since confirmed our results. Our analyses can be found here: https://github.com/cmry/amica/tree/master/reproduction.

Table 8 Overview of different architectures (Arch) their in-domain positive class F_1 scores for Experiment I (T1) and II (T2), the out-of-domain average for D_{all} (*all*), and D_{tox} (*tox*)

Arch	T1						Avg		T2		T3
	D_{twB}	D_{frm}	D_{msp}	D_{ytb}	D_{ask}	D_{twX}	*all*	*tox*	C_{frm}	C_{ask}	D_{tox}
`baseline`	.417	.454	.941	.365	**.561**	**.508**	**.557**	**.389**	**.758**	.579	**.806**
NBSVM	.383	**.486**	.925	**.387**	.476	.396	.551	.385	.703	.604	.797
BiLSTM*	.171	.363	.938	.152	.504	.400	.440	.349	.609	.507	.762
BiLSTM+	.188	.396	**.951**	.160	.438	.341	.417	.337	.541	.505	.737
BiLSTM	.182	.341	.905	.148	.463	.246	.479	.356	.608	.522	.774
CNN*	**.500**	.276	.790	.133	.462	.438	.364	.350	.000	.306	.753
CNN	.444	.416	.816	.000	.498	.438	.464	.342	.000	.610	.754
CNN★	.444	.419	.816	.000	.499	.375	.460	.362	.000	**.647**	.774
C-LSTM*	.000	.421	.875	.095	.000	.000	.449	.329	.094	.425	.757
C-LSTM	.000	.019	.829	.000	.066	.000	.463	.355	.095	.518	.761
C-LSTM★	.000	.057	.853	.075	.008	.000	.278	.358	.296	.506	.756

Baseline model (and scores) is that of Table 4. Reproduction results of Agrawal and Awekar (2018) are denoted by *, their oversampling method by +. Our tuned model versions have no annotation, character level models are denoted by ★

comparitatively balanced class distribution) see much more competitive perfomance compared to the baselines. The same effect can be observed when more data is available; both averages test scores for D_{all} and D_{tox} are comparable to the baseline across almost all architectures. Additionally, the D_{tox} scores indicate that all architectures show about the same overlap on toxicity detection, although interestingly, less so for the neural models than for the baselines.

It can therefore be concluded that the current neural architectures do not provide a solution to the limitations of the task, rather, suffering more in performance. Our experiments do, however, once more illustrate that the proposed techniques of improving the representations of the corpora (by providing more data through merging all sources, and balancing by classifying batches of multiple messages, or conversations) allow the neural models to approach the baseline ballpark. As our goal here was not to completely optimize these architectures, but replication, the proposed techniques still could provide more avenues for further research. Finally, given its robust performance, we will continue to use the baseline model for the next experiment.

6.5 Experiment V

Due to the nature of its experimental set-up (which generates balanced data with simple language use, as shown in Table 2), the crowdsourced data proves easy to classify. Therefore, we do not report out-of-domain averages, as this set would skew them too optimistically, and be uninformative. Regardless, we are primarily

interested in performance when crowdsourced data is added, or used as a replacement for real data. In contrast to the other experiments, the focus will mostly be on the Ask.fm (D_{ask_nl}) and donated (D_{don_nl}) scores (see Table 8). The scores on the Dutch part of the Ask.fm corpus are quite similar to those on the English corpus (.561 vs .598 positive F_1 score), which is in line with earlier results (Van Hee et al. 2018). Moreover, particularly for the small amount of data, the crowdsourced corpus performs surprisingly well on D_{ask_nl} (.516), and significantly better on the donated test data (.667 on D_{don_nl}). This implies that a balanced, controlled bullying set, tailored to the task specifically, does not need a significant amount of data to achieve comparable (or even better) performance, which is a promising result.

Furthermore, in the settings that utilize context representations, training on conversation scopes initially does not seem to improve detection performance in any of the configurations (save for a marginal gain on D_{simC_nl}). However, it does simplify the task in a meaningful way at test-time; whereas a slight gain is obtained for message-level D_{ask_nl} (from .598 F_1-score to .608), when merging both datasets a significant performance boost can be found when training on D_{comb} and testing on D_{askC_nl} (from .264 and .501 to .801 on the combined). Hence, it can be further concluded that enriching the existing training set with crowdsourced data yields meaningful improvements.

Based on these results, we confirm the Experiment II results hold for Dutch: more diverse, larger datasets, and increasing context sizes contributes to better performance on the task. Most importantly, there is enough evidence to support Hypothesis 3: the data generated by the crowdsourcing experiment helps detection rates for our in-the-wild test set, and its combination with externally collected data increases performance with and without additional context. These results underline the potential of this approach to collecting cyberbullying data.

6.6 Suggestions for future work

We hope our experiments have helped to shed light on, and raise further attention to multiple issues with methodological rigor pertaining the task of cyberbullying detection. It is our understanding that the disproportionate amount of work on the (oversimplified) classification task, versus the lack of focus on constructing rich, representative corpora reflecting the actual dynamics of bullying, has made critical assessment of the advances in this task difficult. We would therefore want to particularly stress the importance of simple baselines and the out-of-domain tests that we included in the evaluation criterion for this research. They would provide a fairer comparison for proposed novel classifiers, and a more unified method of evaluation. In line with this, the structural inclusion of domain adaptation techniques seems a logical next step to improve cross-domain performance, specifically those tailored to imbalanced data.

Furthermore, this should be paired with a critical view on the extent to which the full scope of the task is fulfilled. Novel research would benefit from explicitly finding evidence to support its assumptions that classifiers labeled ‘cyberbullying detection’ do more than one-shot, message-level toxicity detection. We would argue that the current framing of the majority of work on the task is still too limited to be

Table 9 Positive class F_1 scores for Experiment IV on Dutch data

Train	T1			T2	
	D_{ask_nl}	D_{sim_nl}	D_{don_nl}	D_{askC_nl}	D_{simC_nl}
D_{ask_nl}	.598	.516	.495	.264	.533
D_{sim_nl}	.273	**.708**	**.667**	.501	.800
D_{comb}	**.608**	.681	.516	**.801**	.808
D_{askC_nl}	.165	.361	.182	.505	.750
D_{simC_nl}	.175	.424	.333	.496	.750
D_{all_nl}	.577	.677	.516	.379	**.821**

Models are fitted on the training proportion of the corpora row-wise and tested column-wise. The best overall test score is noted in bold. The scores of primary interest are highlighted in gray

considered theoretically-defined cyberbullying classification. In our research, we demonstrated several qualitative and quantitative methods that can facilitate such analyses. As popularity of the application of cyberbullying detection grows, this would avoid misrepresenting the conducted work, and that of in-the-wild applications in the future.

We can imagine these conclusions to be relevant for more research within the computational forensics domain: detection of online pedophilia (Bogdanova et al. 2014), aggression and intimidation (Escalante et al. 2017), terrorism and extremism (Ebrahimi et al. 2016; Kaati et al. 2015), and systematic deception Feng et al. (2012)–among others. These are all examples of heavily skewed phenomena residing within more complex networks for which simplification could lead to serious misrepresentation of the task. As with cyberbullying research, a critical evaluation of multiple domains could potentially uncover problematic performance gaps.

While we demonstrated a method of collecting plausible cyberbullying with guaranteed consent, the more valuable sources of real-world data that allow for complex models of social interaction remain restricted. It is our expectation that future modeling will benefit from the construction of much larger (anonymized) corpora—as most fields dealing with language have, and we therefore hope to see future work heading this direction.

7 Conclusion

In this work, we identified several issues that affect the majority of the current research on cyberbullying detection. As it is difficult to collect accurate cyberbullying data in the wild, the field suffers from data scarcity. In an optimal scenario, rich representations capturing all required meta-data to model the complex social dynamics of what the literature defines as cyberbullying would likely prove fruitful. However, one can assume such access to remain restricted for the time

being, and with current social media moving towards private communication, to not be generalizable in the first place. Thus, significant changes need to be made to the empirical practices in this field. To this end, we provided a cross-domain evaluation setup and tested several cyberbullying detection models, under a range of different representations to potentially overcome the limitations of the available data, and provide a fair, rigorous framework to facilitate direct model comparison for this task.

Additionally, we formed three hypotheses we would expect to find evidence for during these evaluations: (1) the corpora are too small and heterogeneous to represent the strong variation in language use for both bullying and neutral content across platforms accurately, (2) the positive instances are biased, predominantly capturing toxicity, and no other dimensions of bullying, and finally (3) crowdsourcing poses a resource to generate plausible cyberbullying events, and that can help expand the available data and improve the current models.

We found evidence for all three hypotheses: previous cyberbullying models generalize poorly across domains, simple BoW baselines prove difficult to improve upon, there is considerable overlap between toxicity classification and cyberbullying detection, and crowdsourced data yields well-performing cyberbullying detection models. We believe that the results of Hypotheses 1 and 2 in particular are principal hurdles that need to be tackled to advance this field of research. Furthermore, we showed that both leveraging training data from all openly available corpora, and shifting representations to include context meaningfully improves performance on the overall task. Therefore, we believe both should be considered as an evaluation point in future work. More so given that we show that these do not solve the existing limitations of the currently available corpora, and could therefore provide avenues for future research focusing on collecting (richer) data. Lastly, we show reproducibility of models that previously demonstrated state-of-the-art performance on this task to fail. We hope that the observations and contributions made in this paper can aid to improve rigor in future cyberbullying detection work.

Acknowledgements We would like to thank Prodromos Ninas, Kostas Stoitsas, and Alejandra Hernández Rejón for carrying out a range of trial experiments on this task, Bram Willemsen for helpful remarks, and in particular Ákos Kádár for the many discussions throughout all stages of the current work. Finally, we want to express our gratitude to all authors providing cyberbullying data open-source. Most of all to Agrawal et al. for making their work reproducible, and thereby facilitating rigorous evaluations in this task.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org/. Software available from tensorflow.org.

Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 141–153). Cham: Springer International Publishing.

Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, *15*(11), 937–946.

Bayzick, J., Kontostathis, A., & Edwards, L. (2011). Detecting the presence of cyberbullying using computer software. In *Proceedings of the 3rd international web science conference*. WebSci11; 2011.

Beran, T., & Li, Q. (2008). The relationship between cyberbullying and school bullying. *The Journal of Student Wellbeing*, *1*(2), 16–33.

Bessiere, K., Kiesler, S., Kraut, R., & Boneva, B. S. (2008). Effects of internet use and social resources on changes in depression. *Information, Community & Society*, *11*(1), 47–70.

Binns, A. (2013). Facebook's ugly sisters: Anonymity and abuse on formspring and ask. fm. Media Education Research Journal

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 440–447).

Bogdanova, D., Rosso, P., & Solorio, T. (2014). Exploring high-level features for detecting cyberpedophilia. *Computer Speech & Language*, *28*(1), 108–120.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bretschneider, U., Wöhner, T., & Peters, R. (2014). Detecting Online Harassment in Social Networks. In *35th International Conference on Information Systems* Li 2007 (pp. 1–14).

Broadcasting Standards Authority: What not to swear: The acceptability of words in broadcasting. Retrieved March 03, 2016 from, http://bsa.govt.nz/images/assets/Research/Acceptability_of_Words_2013_WEB.pdf (2013).

Broeck, E., Poels, K., Vandebosch, H., & Royen, K. (2014). Online perspective-taking as an intervention tool against cyberbullying. Annual review of cybertherapy and telemedicine (pp. 113–117).

Chan, Y.S., & Ng, H.T. (2006). Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 89–96). Association for Computational Linguistics.

Chen, M., Xu, Z., Weinberger, K., & Sha, F. (2012). Marginalized denoising autoencoders for domain adaptation. arXiv preprint arXiv:1206.4683.

Cheng, L., Guo, R., Candan, K.S., & Liu, H. (2020). Representation learning for imbalanced cross-domain classification. In *Proceedings of the 2020 SIAM international conference on data mining*. (pp. 478–486). SIAM.

Cheng, L., Guo, R., Silva, Y., Hall, D., & Liu, H. (2019). Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM international conference on data mining* (pp. 235–243). SIAM.

Chollet, F., et al. (2015). Keras. https://keras.io.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018.

Dadvar, M. (2014). Experts and machines united against cyberbullying. Ph.D. thesis, University of Twente, Netherlands. https://doi.org/10.3990/1.9789036537391.

Dadvar, M., Trieschnigg, D., de Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian conference on artificial intelligence* (pp. 275–281). Springer.

Daumé III, H. (2009). Frustratingly easy domain adaptation. arXiv preprint arXiv:0907.1815.

DeHue, F., Bolman, C., & Völlink, T. (2008). Cyberbullying: Youngsters' experiences and parental perception. *CyberPsychology & Behavior*, *11*(2), 217–223.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. the social mobile web. In *5th international AAAI conference on weblogs and social media, Barcelona, Catalonia, Spain*.

Duggan, M. (2015). *Mobile messaging and social media 2015*. Washington, D.C: Pew Research Center.

Ebrahimi, M., Suen, C. Y., & Ormandjieva, O. (2016). Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, *18*, 33–49.

Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y Gómez, M., & Villaseñor-Pineda, L. (2017). Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, *89*, 99–111.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning*, *9*(2008), 1871–1874. https://doi.org/10.1038/oby.2011.351. http://dl.acm.org/citation.cfm?id=1442794.

Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (Volume 2: Short Papers) (pp. 171–175).

Finkel, J.R., & Manning, C.D. (2009). Hierarchical bayesian domain adaptation. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics* (pp. 602–610).

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with lstm recurrent networks. *Journal of Machine Learning Research*, *3*, 115–143.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach.

Hee, C.V., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., Pauw, G.D., & Daelemans, W. (2015). Detection and Fine-Grained Classification of Cyberbullying Events. In *International conference recent advances in natural language processing (RANLP)* (pp. 672–680).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q., Mishra, S. (2015). Prediction of cyberbullying incidents on the Instagram social network. In MobiSys, (p. 2014). arxiv:1503.03909.

Jiang, J., & Zhai, C. (2007). A two-stage approach to domain adaptation for statistical classifiers. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 401–410).

Kaati, L., Omer, E., Prucha, N., & Shrestha, A. (2015). Detecting multipliers of jihadism on twitter. In *2015 IEEE international conference on data mining workshop (ICDMW)* (pp. 954–960). IEEE.

Kann, L., Kinchen, S., Shanklin, S. L., Flint, K. H., Kawkins, J., Harris, W. A., et al. (2014). Youth risk behavior surveillance-united states, 2013. *Morbidity and Mortality Weekly Report: Surveillance Summaries*, *63*(Suppl 4), 1–168.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. Text mining: Applications and theory. John Wiley & Sons, Ltd, Chichester, UK (pp. 149–164).

Kontostathis, A., Reynolds, K. (2013). Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 3rd annual ACM web science conference*, WebSci 2013 (2013) (pp. 195–204).

Kowalski, R. M., Limber, S. P., & Agatston, P. W. (2012). *Cyberbullying: Bullying in the digital age*. John Wiley & Sons.

Lenhart, A., Madden, M., Smith, A., Purcell, K., Zickuhr, K., & Rainie, L. (2011). Teens, kindness and cruelty on social network sites: How american teens navigate the new world of. Pew Internet & American Life Project.

Li, S., Ju, S., Zhou, G., & Li, X. (2012). Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 139–148). Association for Computational Linguistics.

Livingstone, S., Haddon, L., Görzig, A., & Ólafsson, K. (2011). EU Kids Online II: Final Report 2011. Tech. rep., EU Kids Online. http://www.lse.ac.uk/collections/EUKidsOnline/ (p. 293).

Ma, J., Zhang, Y., & Zhu, J. (2014). Tagging the web: Building a robust web tagger with neural network. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (Vol. 1: Long Papers) (pp. 144–154).

Madden, M., Lenhart, A., Cortesi, S. (2013). Teens, social media, and privacy. Tech. rep., Pew Internet. http://www.lateledipenelope.it/public/52dff2e35b812.pdf.

McClosky, D., Charniak, E., Johnson, M.: Reranking and self-training for parser adaptation. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 337–344). Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Millwood-Hargreave, A. (2000). *Delete expletives? research undertaken jointly by the advertising standards authority, british broadcasting corporation, broadcasting standards commission and the independent television commission*. London: ASA, BBC, BSC and ITC.

Nahar, V., Li, X., & Pang, C. (2013). An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, *3*(May), 238–247.

Noswearing.com: Bad Word List & Swear Filter. Retrieved March 03, 2016 from, http://www.noswearing.com (2016).

Ólafsson, K., Livingstone, S., & Haddon, L. (2013). *Children's Use of Online Technologies in Europe: A review of the European evidence base*. EU Kids Online: Tech. Rep. May.

Pan, S.J., Ni, X., Sun, J.T., Yang, Q., & Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web* (pp. 751–760).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). http://www.aclweb.org/anthology/D14-1162.

Privitera, C., & Campbell, M. A. (2009). Cyberbullying: The new face of workplace bullying? *CyberPsychology & Behavior*, *12*(4), 395–400.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). ELRA, Valletta, Malta (2010). http://is.muni.cz/publication/884893/en.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. In *10th international conference on machine learning and applications and workshops (2011)* (pp. 241–244).

Robers, S., Zhang, A., & Morgan, R.E. (2015). Indicators of school crime and safety: 2014. nces 2015-072/ncj 248036. National Center for Education Statistics.

Rosa, H., Carvalho, J.P., Calado, P., Martins, B., Ribeiro, R., & Coheur, L. (2018). Using fuzzy fingerprints for cyberbullying detection in social networks. In *2018 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 1–7). IEEE.

Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J.P. (2018). A "deeper" look at detecting cyberbullying in social networks. In *2018 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., et al. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, *93*, 333–345.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *LREC*, (pp. 486–493).

Schnabel, T., & Schütze, H. (2014). Flors: Fast and simple domain adaptation for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, *2*, 15–26.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

Sharp, S., & Smith, P. K. (2002). *School bullying: Insights and perspectives*. : Routledge.

Smith, A., & Anderson, M. (2018). *Social media use in 2018*. Washington, D.C.: Pew Research Center.

Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015). Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015* (pp. 280–285). ACM.

Steijn, W. M., & Schouten, A. P. (2013). Information sharing and relationships on social networking sites. *Cyberpsychology, Behavior, and Social Networking*, *16*(8), 582–587.

Sui, J. (2015). Understanding and fighting bullying with machine learning. Ph.D. thesis, The University of Wisconsin-Madison.

Sun, C., Qiu, X., Xu, Y., & Huang, X.(2019). How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics* (pp. 194–206). Springer.

Thain, N., Dixon, L., & Wulczyn, E. (2017). *Wikipedia Talk Labels: Toxicity*,. https://doi.org/10.6084/m9.figshare.4563973.v2. https://figshare.com/articles/Wikipedia_Talk_Labels_Toxicity/4563973.

Tomkins, S., Getoor, L., Chen, Y., & Zhang, Y. (2018). A socio-linguistic model for cyberbullying detection. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 53–60). IEEE.

Tulkens, S., Emmery, C., & Daelemans, W. (2016). Evaluating unsupervised dutch word embeddings as a linguistic resource. In: N.C.C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Vaez, M., Ekberg, K., & LaFlamme, L. (2004). Abusive events at work among young working adults: Magnitude of the problem and its effect on self-rated health. Relations industrielles/industrial relations (pp. 569–584).

Valkenburg, P. M., & Peter, J. (2007). Preadolescents' and adolescents' online communication and their closeness to friends. *Developmental Psychology*, *43*(2), 267.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., et al. (2018). Automatic detection of cyberbullying in social media text. *PLoS ONE*, *13*(10), 1–22. https://doi.org/10.1371/journal.pone.0203794.

Van Hee, C., Verhoeven, B., Lefever, E., De Pauw, G., Daelemans, W., & Hoste, V. (2015). Guidelines for the fine-grained analysis of cyberbullying. Tech. rep., version 1.0. Technical Report LT3 15-01, LT3, Language and Translation Technology Team—Ghent University.

Vilain, M., Su, J., Lubar, S. (2007). Entity extraction is a boring solved problem—or is it? In *Human language technologies 2007: The conference of the North American chapter of the Association for Computational Linguistics*; Companion Volume, Short Papers (pp. 181–184).

Wang, S., & Manning, C.D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2* (pp. 90–94). Association for Computational Linguistics.

Willard, N. E. (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. : Research Press.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on World Wide Web* (pp. 1391–1399).

Xu, J., Jun, K., Zhu, X., & Bellmore, A. (2012). Learning from Bullying Traces in Social Media. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 656–666). Association for Computational Linguistics.

Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., & Edwards, L. (2009). Detection of Harassment on Web 2.0. In *Proceedings of the content analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*.

Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J.P., Kowalski, R., Hu, H., Luo, F., Macbeth, J., & Dillon, E. (2016). Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)* (pp. 740–745). IEEE.

Zhao, R., & Mao, K. (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, *8*(3), 328–339.

Zhao, R., Zhou, A., & Mao, K. (2016): Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking, ICDCN '16* (pp. 43:1–43:6). ACM, New York, NY, USA. https://doi.org/10.1145/2833312.2849567.

Zhou, C., Sun, C., Liu, Z., & Lau, F. (2015). A c-lstm neural network for text classification. arXiv preprint arXiv:1511.08630.