Understanding Abuse: A Typology of Abusive Language Detection Subtasks

Zeerak Waseem
Department of Computer Science
University of Sheffield
United Kingdom
z.w.butt@sheffield.ac.uk

Thomas Davidson
Department of Sociology
Cornell University
Ithica, NY
trd54@cornell.edu

Dana Warmsley
Department for Applied Mathematics
Cornell University
Ithica, NY
dw457@cornell.edu

Ingmar Weber
Qatar Computing Research Institute
HBKU
Doha, Qatar
iweber@hbku.edu.qa

Abstract

As the body of research on abusive language detection and analysis grows, there is a need for critical consideration of the relationships between different subtasks that have been grouped under this label. Based on work on hate speech, cyberbullying, and online abuse we propose a typology that captures central similarities and differences between subtasks and we discuss its implications for data annotation and feature construction. We emphasize the practical actions that can be taken by researchers to best approach their abusive language detection subtask of interest.

1 Introduction

There has been a surge in interest in the detection of abusive language, hate speech, cyberbullying, and trolling in the past several years (Schmidt and Wiegand, 2017). Social media sites have also come under increasing pressure to tackle these issues. Similarities between these subtasks have led scholars to group them together under the umbrella terms of "abusive language", "harmful speech", and "hate speech" (Nobata et al., 2016; Faris et al., 2016; Schmidt and Wiegand, 2017) but little work has been done to examine the relationship between them. As each of these subtasks seeks to address a specific yet partially overlapping phenomenon, we believe that there is much to gain by studying how they are related.

The overlap between subtasks is illustrated by the variety of labels used in prior work. For example, in annotating for cyberbullying events, Van Hee et al. (2015b) identifies discriminative remarks (racist, sexist) as a subset of "insults", whereas Nobata et al. (2016) classifies similar remarks as "hate speech" or "derogatory language". Waseem and Hovy (2016) only consider "hate speech" without regard to any potential overlap with bullying or otherwise offensive language, while Davidson et al. (2017) distinguish hate speech from generally offensive language. Wulczyn et al. (2017) annotates for personal attacks, which likely encompasses identifying cyberbullying, hate speech, and offensive language. The lack of consensus has resulted in contradictory annotation guidelines - some messages considered as hate speech by Waseem and Hovy (2016) are only considered derogatory and offensive by Nobata et al. (2016) and Davidson et al. (2017).

To help to bring together these literatures and to avoid these contradictions, we propose a typology that synthesizes these different subtasks. We argue that the differences between subtasks within abusive language can be reduced to two primary factors:

1. *Is the language directed towards a specific individual or entity or is it directed towards a generalized group?*

2. *Is the abusive content explicit or implicit?*

Each of the different subtasks related to abu-

sive language occupies one or more segments of this typology. Our aim is to clarify the similarities and differences between subtasks in abusive language detection to help researchers select appropriate strategies for data annotation and modeling.

2 A typology of abusive language

Much of the work on abusive language subtasks can be synthesized in a two-fold typology that considers whether (i) the abuse is directed at a specific target, and (ii) the degree to which it is explicit.

Starting with the targets, abuse can either be directed towards a specific individual or entity, or it can be used towards a generalized *Other*, for example people with a certain ethnicity or sexual orientation. This is an important sociological distinction as the latter references a whole category of people rather than a specific individual, group, or organization (see Brubaker 2004, Wimmer 2013) and, as we discuss below, entails a linguistic distinction that can be productively used by researchers. To better illustrate this, the first row of Table 1 shows examples from the literature of directed abuse, where someone is either mentioned by name, tagged by a username, or referenced by a pronoun.[1] Cyberbullying and trolling are instances of directed abuse, aimed at individuals and online communities respectively. The second row shows cases with abusive expressions towards generalized groups such as racial categories and sexual orientations. Previous work has identified instances of hate speech that are both directed and generalized (Burnap and Williams, 2015; Waseem and Hovy, 2016; Davidson et al., 2017), although Nobata et al. (2016) come closest to making a distinction between directed and generalized hate.

The other dimension is the extent to which abusive language is explicit or implicit. This is roughly analogous to the distinction in linguistics and semiotics between *denotation*, the literal meaning of a term or symbol, and *connotation*, its sociocultural associations, famously articulated by Barthes (1957). Explicit abusive language is that which is unambiguous in its *potential* to be abusive, for example language that contains racial or homophobic slurs. Previous research has indicated a great deal of variation within such language (Warner and Hirschberg, 2012; Davidson et al., 2017), with abusive terms being used in a colloquial manner or by people who are victims of abuse. Implicit abusive language is that which does not immediately imply or denote abuse. Here, the true nature is often obscured by the use of ambiguous terms, sarcasm, lack of profanity or hateful terms, and other means, generally making it more difficult to detect by both annotators and machine learning approaches (Dinakar et al., 2011; Dadvar et al., 2013; Justo et al., 2014). Social scientists and activists have recently been paying more attention to implicit, and even unconscious, instances of abuse that have been termed "micro-aggressions" (Sue et al., 2007). As the examples show, such language may nonetheless have extremely abusive connotations. The first column of Table 1 shows instances of explicit abuse, where it should be apparent to the reader that the content is abusive. The messages in the second column are implicit and it is harder to determine whether they are abusive without knowing the context. For example, the word "them" in the first two examples in the generalized and implicit cell refers to an ethnic group, and the words "skypes" and "Google" are used as euphemisms for slurs about Jews and African-Americans respectively. Abuse using sarcasm can be even more elusive for detection systems, for instance the seemingly harmless comment praising someone's intelligence was a sarcastic response to a beauty pageant contestants unsatisfactory answer to a question (Dinakar et al., 2011).

3 Implications for future research

In the following section we outline the implications of this typology, highlighting where the existing literatures indicate how we can understand, measure, and model each subtype of abuse.

3.1 Implications for annotation

In the task of annotating documents that contain bullying, it appears that there is a common understanding of what cyberbullying entails: an intentionally harmful electronic attack by an individual or group against a victim, usually repetitive in nature (Dadvar et al., 2013). This consensus allows for a relatively consistent set of annotation guidelines across studies, most of which simply ask annotators to determine if a post contains bullying or harassment (Dadvar et al., 2014; Kontostathis et al., 2013; Bretschneider et al., 2014).

[1] All punctuation is as reported in original papers. We have added all the * symbols.

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" (Van Hee et al., 2015a), "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" (Davidson et al., 2017), "Youre one of the ugliest b*tches Ive ever fucking seen" (Kontostathis et al., 2013).	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" (Dinakar et al., 2012), "((((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" (Hine et al., 2017), "you're intelligence is so breathtaking!!!!!!" (Dinakar et al., 2011)
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" (Nobata et al., 2016), "So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!" (Kwok and Wang, 2013).	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap and Williams, 2015), "most of them come north and are good at just mowing lawns" (Dinakar et al., 2011), "Gas the skypes" (Magu et al., 2017)

Table 1: **Typology of abusive language.**

High inter-annotator agreement on cyberbullying tasks (93%) (Dadvar et al., 2013) further indicates a general consensus around the features of cyberbullying (Van Hee et al., 2015b). After bullying has been identified annotators are typically asked more detailed questions about the extremity of the bullying, the identification of phrases that indicate bullying, and the roles of users as bully/victim (Dadvar et al., 2014; Van Hee et al., 2015b; Kontostathis et al., 2013).

We expect that consensus may be due to the directed nature of the phenomenon. Cyberbullying involves a victim whom annotators can identify and relatively easily discern whether statements directed towards the victim should be considered abusive. In contrast, in work on annotating harassment, offensive language, and hate speech there appears to be little consensus on definitions and lower inter-annotator agreement ($\kappa \approx 0.60-0.80$) (Ross et al., 2016; Waseem, 2016a; Tulkens et al., 2016; Bretschneider and Peters, 2017) are obtained. Given that these tasks are often broadly defined and the target is often generalized, all else being equal, it is more difficult for annotators to determine whether statements should be considered abusive. Future work in these subtasks should aim to have annotators distinguish between targeted and generalized abuse so that each subtype can be modeled more effectively.

Annotation (via crowd-sourcing and other methods) tends to be more straightforward when explicit instances of abusive language can be identified and agreed upon (Waseem, 2016b), but is considerably more difficult when implicit abuse is considered (Dadvar et al., 2013; Justo et al., 2014; Dinakar et al., 2011). The connotations of language can be difficult to classify without domain-specific knowledge. Furthermore, while some argue that detailed guidelines can help annotators to make more subtle distinctions (Davidson et al., 2017), others find that they do not improve the reliability of non-expert classifications (Ross et al., 2016). In such cases, expert annotators with domain specific knowledge are preferred as they tend to produce more accurate classifications (Waseem, 2016a).

Ultimately, the nature of abusive language can be extremely subjective, and researchers must endeavor to take this into account when using human annotators. Davidson et al. (2017), for instance, show that annotators tend to code racism as hate speech at a higher rate than sexism. As such, it is important that researchers consider the social biases that may lead people to disregard certain types of abuse.

The type of abuse that researchers are seeking to identify should guide the annotation strategy. Where subtasks occupy multiple cells in our typology, annotators should be allowed to make nuanced distinctions that differentiate between different types of abuse. In highlighting the major differences between different abusive language detection subtasks, our typology indicates that different annotation strategies are appropriate depending on the type of abuse.

3.2 Implications for modeling

Existing research on abusive language online has used a diverse set of features. Moving forward, it is important that researchers clarify which features are most useful for which subtasks and which subtasks present the greatest challenges. We do not attempt to review all the features used (see Schmidt and Wiegand 2017 for a detailed review)

but make suggestions for which features could be most helpful for the different subtasks. For each aspect of the typology, we suggest features that have been shown to be successful predictors in prior work. Many features occur in more than one form of abuse. As such, we do not propose that particular features are necessarily unique to each phenomenon, rather that they provide different insights and should be employed depending on what the researcher is attempting to measure.

Directed abuse. Features that help to identify the target of abuse are crucial to directed abuse detection. Mentions, proper nouns, named entities, and co-reference resolution can all be used in different contexts to identify targets. Bretschneider and Peters (2017) use a multi-tiered system, first identifying offensive statements, then their severity, and finally the target. Syntactical features have also proven to be successful in identifying abusive language. A number of studies on hate speech use part-of-speech sequences to model the expression of hatred (Warner and Hirschberg, 2012; Gitari et al., 2015; Davidson et al., 2017). Typed dependencies offer a more sophisticated way to capture the relationship between terms (Burnap and Williams, 2015). Overall, there are many tools that researchers can use to model the relationship between abusive language and targets, although many of these require high-quality annotations to use as training data.

Generalized abuse. Generalized abuse online tends to target people belonging to a small set of categories, primarily racial, religious, and sexual minorities (Silva et al., 2016). Researchers should consider identifying forms of abuse unique to each target group addressed, as vocabularies may depend on the groups targeted. For example, the language used to abuse trans-people and that used against Latin American people are likely to differ, both in the nouns used to denote the target group and the other terms associated with them. In some cases a lexical method may therefore be an appropriate strategy. Further research is necessary to determine if there are underlying syntactic structures associated with generalized abusive language.

Explicit abuse Explicit abuse, whether directed or generalized, is often indicated by specific keywords. Hence, dictionary-based approaches may be well suited to identify this type of abuse (Warner and Hirschberg, 2012; Nobata et al., 2016), although the presence of particular words should not be the only criteria, even terms that denote abuse may be used in a variety of different ways (Kwok and Wang, 2013; Davidson et al., 2017). Negative polarity and sentiment of the text are also likely indicators of explicit abuse that can be leveraged by researchers (Gitari et al., 2015).

Implicit abuse. Building a specific lexicon may prove impractical, as in the case of the appropriation of the term "skype" in some forums (Magu et al., 2017). Still, even partial lexicons may be used as seeds to inductively discover other keywords by use of a semi-supervised method proposed by King et al. (2017). Additionally, character n-grams have been shown to be apt for abusive language tasks due to their ability to capture variation of words associated with abuse (Nobata et al., 2016; Waseem, 2016a). Word embeddings are also promising ways to capture terms associated with abuse (Djuric et al., 2015; Badjatiya et al., 2017), although they may still be insufficient for cases like 4Chan's connotation of "skype" where a word has a dominant meaning and a more subversive one. Furthermore, as some of the above examples show, implicit abuse often takes on complex linguistic forms like sarcasm, metonymy, and humor. Without high quality labeled data to learn these representations, it may be difficult for researchers to come up with models of syntactic structure that can help to identify implicit abuse. To overcome these limitations researchers may find it prudent to incorporate features beyond just textual analysis, including the characteristics of the individuals involved (Dadvar et al., 2013) and other extra-textual features.

4 Discussion

This typology has a number of implications for future work in the area.

First, we want to encourage researchers working on these subtasks to learn from advances in other areas. Researchers working on purportedly distinct subtasks are often working on the same problems in parallel. For example, the field of hate speech detection can be strengthened by interactions with work on cyberbullying, and vice versa, since a large part of both subtasks consists of identifying targeted abuse.

Second, we aim to highlight the important distinctions within subtasks that have hitherto been ignored. For example, in much hate speech research, diverse types of abuse have been lumped

together under a single label, forcing models to account for a large amount of within-class variation. We suggest that fine-grained distinctions along the axes allows for more focused systems that may be more effective at identifying particular types of abuse.

Third, we call for closer consideration of how annotation guidelines are related to the phenomenon of interest. The type of annotation and even the choice of annotators should be motivated by the nature of the abuse. Further, we welcome discussion of annotation guidelines and the annotation process in published work. Many existing studies only tangentially mention these, sometimes never explaining how the data were annotated.

Fourth, we encourage researchers to consider which features are most appropriate for each subtask. Prior work has found a diverse array of features to be useful in understanding and identifying abuse, but we argue that different feature sets will be relevant to different subtasks. Future work should aim to build a more robust understanding of when to use which types of features.

Fifth, it is important to emphasize that not all abuse is equal, both in terms of its effects and its detection. We expect that social media and website operators will be more interested in identifying and dealing with explicit abuse, while activists, campaigners, and journalists may have more incentive to also identify implicit abuse. Targeted abuse such as cyberbullying may be more likely to be reported by victims and thus acted upon than generalized abuse. We also expect that implicit abuse will be more difficult to detect and model, although methodological advances may make such tasks more feasible.

5 Conclusion

We have presented a typology that synthesizes the different subtasks in abusive language detection. Our aim is to bring together findings in these different areas and to clarify the key aspects of abusive language detection. There are important analytical distinctions that have been largely overlooked in prior work and through acknowledging these and their implications we hope to improve abuse detection systems and our understanding of abusive language.

Rather than attempting to resolve the "definitional quagmire" (Faris et al., 2016) involved in neatly bounding and defining each subtask we encourage researchers to think carefully about the phenomena they want to measure and the appropriate research design. We intend for our typology to be used both at the stage of data collection and annotation and the stage of feature creation and modeling. We hope that future work will be more transparent in discussing the annotation and modeling strategies used, and will closely examine the similarities and differences between these subtasks through empirical analyses.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. pages 759–760.

Roland Barthes. 1957. *Mythologies*. Seuil.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Uwe Bretschneider, Thomas Whner, and Ralf Peters. 2014. Detecting online harassment in social networks. In *ICIS 2014 Proceedings: Conference Theme Track: Building a Better World through IS*.

Rogers Brubaker. 2004. *Ethnicity without groups*. Harvard University Press.

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

Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and machines against bullies: a hybrid approach to detect cyberbullies. In *Conference on Artificial Intelligence*. Springer International Publishing.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*. Springer, pages 693–696.

Thomas Davidson, Dana Warmsley, Micheel Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media*. Montreal, Canada, pages 512–515.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and

mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3):18.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. *The Social Mobile Web* 11(02).

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

Robert Faris, Amar Ashar, Urs Gasser, and Daisy Joo. 2016. Understanding harmful speech online. *Berkman Klein Center Research Publication* 21.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* 10(4):215–230.

Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. A longitudinal measurement study of 4chan's politically incorrect forum and its effect on the web. In *Proceedings of the Eleventh International Conference on Web and Social Media*. Montreal, Canada, pages 92–101.

Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Ins Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems* 69:124 – 133.

Gary King, Patrick Lam, and Margaret E Roberts. 2017. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science* .

April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, New York, NY, USA, WebSci '13, pages 195–204.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI'13, pages 1621–1622.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *Proceedings of the Eleventh International Conference on Web and Social Media*. Montreal, Canada, pages 608–612.

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

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*. pages 6–9.

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

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media*. Cologne, Germany, pages 687–690.

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *American Psychologist* 62(4):271–286.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. The automated detection of racist discourse in dutch social media. *CLIN Journal* 6:3–20.

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

Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Véronique Hoste, and Walter Daelemans. 2015b. Guidelines for the fine-grained analysis of cyberbullying. Technical report, LT3, Ghent University, Belgium.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, LSM '12, pages 19–26.

Zeerak Waseem. 2016a. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, pages 138–142.

Zeerak Waseem. 2016b. *Automatic Hate Speech Detection*. Master's thesis, University of Copenhagen.

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

NAACL Student Research Workshop. Association for Computational Linguistics, San Diego, California, pages 88–93.

Andreas Wimmer. 2013. *Ethnic boundary making: Institutions, power, networks*. Oxford University Press.

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