Modeling Users and Online Communities for Abuse Detection: A Position on Ethics and Explainability

Pushkar Mishra[★], **Helen Yannakoudakis**[♠], **Ekaterina Shutova**[♣]
[★] Facebook AI, London, United Kingdom
[♠] Department of Informatics, King's College London, United Kingdom
[♣] Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
pushkarmishra@fb.com, helen.yannakoudakis@kcl.ac.uk, e.shutova@uva.nl

Abstract

Abuse on the Internet is an important societal problem of our time. Millions of Internet users face harassment, racism, personal attacks, and other types of abuse across various platforms. The psychological effects of abuse on individuals can be profound and lasting. Consequently, over the past few years, there has been a substantial research effort towards automated abusive language detection in the field of NLP. In this position paper, we discuss the role that modeling of users and online communities plays in abuse detection. Specifically, we review and analyze the state of the art methods that leverage user or community information to enhance the understanding and detection of abusive language. We then explore the ethical challenges of incorporating user and community information, laying out considerations to guide future research. Finally, we address the topic of explainability in abusive language detection, proposing properties that an explainable method should aim to exhibit. We describe how user and community information can facilitate the realization of these properties and discuss the effective operationalization of explainability in view of the properties.

1 Introduction

With the advent of social media, anti-social and abusive behavior has become a prominent occurrence online. Undesirable psychological effects of abuse on individuals make it an important societal problem of our time. Munro (2011) studied the ill-effects of online abuse on children, concluding that children may develop depression, anxiety, and other mental health problems as a result of their encounters online. *Pew Research Center*, in its latest report on online harassment (Duggan, 2017), revealed that 40% of adults in the United States have experienced abusive behavior online, of which 18% have faced severe forms of harassment, e.g., that of sexual nature. These statistics stress the need for automated detection and moderation systems. Hence, in recent years, a new research effort on abusive language detection has sprung up in NLP.

That said, the notion of abuse has proven elusive and difficult to formalize. Different norms across different (online) platforms can affect what is considered abusive (Chandrasekharan et al., 2018). In the context of natural language, *abuse* is a term that encompasses many different fine-grained types of negative expressions. For example, Nobata et al. (2016) use it to collectively refer to hate speech, derogatory language and profanity, while Mishra et al. (2018a) use it to discuss racism and sexism. The definitions for different types of abuse tend to be overlapping and ambiguous. However, regardless of the specific type, we define abuse as *any expression that is meant to denigrate or offend a particular person or group*. Taking a course-grained view, Waseem et al. (2017) classify abuse into broad categories based on *explicitness* and *directness*. *Explicit* abuse comes in the form of expletives, derogatory words or threats, while *implicit* abuse has a more subtle appearance characterized by the presence of ambiguous terms and figures of speech such as metaphor or sarcasm. *Directed* abuse targets a particular individual as opposed to *generalized* abuse which is aimed at a larger group such as a particular gender or ethnicity.

To date, several approaches to automated detection of abusive language have been proposed, including rule-based (Spertus, 1997; Razavi et al., 2010; Wiegand et al., 2018), linguistic and social feature engineering (Yin et al., 2009; Sood et al., 2012; Warner and Hirschberg, 2012; Salminen et al., 2018), utilizing distributed representations from neural networks (Djuric et al., 2015; Mehdad and Tetreault, 2016; Nobata et al., 2016) or applying deep neural networks directly (Park and Fung, 2017; Pavlopoulos et al., 2017a; Mishra et al., 2018a). Researchers have also explored multi-task learning settings with objectives such as emotion

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3374–3385
November 7–11, 2021. ©2021 Association for Computational Linguistics

detection (Rajamanickam et al., 2020; Samghabadi et al., 2019). We refer the reader to recent surveys of the field (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018) for a detailed literature review.

More recently, researchers have noted that the linguistic features of a comment alone may not be sufficient to classify it as abusive or not. Information of the user who posted the comment, and of the surrounding social community of that user, further provides valuable insights into the abusiveness of the comment. An example of this is the study by Zook (2012), which mapped the locations of racist tweets in response to President Obama's re-election to show that such tweets were not uniformly distributed across the United States but instead came from specific geographical communities of users. Other works have also shown how users on online platforms organize into communities based on factors such as shared beliefs, stereotypes, linguistic norms, or geographical propinquity (Jurgens, 2013; Nguyen and Rosé, 2011).

In this paper, we focus on the role that modeling of users and communities plays in the automated detection of abusive language on online platforms. Specifically, we investigate the different state of the art methods that leverage user or community information to enhance the understanding and detection of abusive language. While these methods have yielded high performance gains, there has been little discussion of the kinds of information they capture. We provide a comprehensive review of these methods, analyzing the information they encode about users or communities and the relevance of that for detection of abusive language. We then explore the ethical considerations of incorporating user and community information in such methods, providing guidance for future research. Finally, we address the topic of explainability in abusive language detection, proposing properties that an explainable detection method should aim to exhibit. We describe how user and community information can facilitate the realization of these properties and discuss the effective operationalization of explainability in view of the properties.

2 Why the user and community matter

Throughout the paper, *user* refers to the user of an online platform who may have posted a comment that is to be classified as abusive or not. The *community* of this user comprises other users and contents that they interact with on the online platform. In other words, community refers to the neighborhood of the user in the social graph of the platform. Conversations online are inherently contextual. Consequently, abuse on online platforms can only be effectively interpreted within a larger *context* (Gao and Huang, 2017) rather than in isolation. This is especially true for implicit or generalized abuse, which are harder to interpret than explicit abuse for humans and machines alike. Information of the user who posted the comment, or of the surrounding community including the targets of the comment, offers insights into several aspects of the context that are otherwise not accessible through the linguistic content of the comment alone. Here, *information* may refer to demographic traits like age or gender, knowledge about linguistic behavior, location details, etc. Below we categorize and discuss the aspects of the context relevant to abusive language detection.

Sociolinguistic norms. *Sociolinguistics* studies the effects of society on language and its usage. Researchers in the past have explored the links between the structures and norms of real-world communities and the linguistic practices of people (D'Arcy and Young, 2012). As in the physical world, individuals and communities on online platforms also abide by certain norms, which may be guided by their cultural backgrounds and/or are based on the standards laid down by the platforms themselves. These norms and standards reflect expectations of *respectful* behavior, local customs and language patterns within a region, etc. (Ben-David and Fernández, 2016). Consequently, the decision of what is considered abusive must be made taking into account the sociolinguistic norms. User and community information, when leveraged alongside linguistic features, helps capture the relevant sociolinguistic norms in a myriad of ways. For example, a comment may contain the *n*-word, but interpretation of its use and or the intent is greatly facilitated by the knowledge of the ethnicity of the user who wrote the comment and/or the ethnicity of the target user or community.

Linguistic variations. Another aspect comes from looking at implicit abuse, whereby a user may utilize novel *slangs* or conventional words in unconventional ways, e.g., as a racial slur or as a name for some specific demographic (Waseem et al., 2017). Information about how a term is being used by other members of a user's community, e.g., in abusive contexts or otherwise, can help decipher linguistic variations that come up from time to time. In fact, it is usually the users with strong ties who are responsible for popularizing language variations as well as for spreading hate speech (Del Tredici and Fernández, 2018; Ribeiro et al., 2018). Therefore, having user and community information alongside linguistic features helps capture linguistic variations and their diffusion.

Prevailing stereotypes. Previous research has shown that prevailing stereotypes often form the basis and justification of abuse. For example, many twitter accounts were open about their anger and hatred for Muslims in the wake of the Rochdale scandal that involved several British–Asian men getting convicted for child grooming (Awan, 2014). Stereotypes are not only explicit but implicit too (Hinton, 2017), which often show up as implicit and subtle abuse in the form of sarcasm, racist jokes, or unnecessary associations. While explicit stereotypes are consciously endorsed, and may be controllable, implicit stereotypes are thought to be shaped by experience and based on learned associations (Byrd, 2019). User and community information plays an important role in the identification of such stereotypes. For example, if the location of users is available alongside linguistic features of the comments they post, one can quickly discover the presence (or absence) of correlations between specific regions and specific kinds of abuse. Moreover, shared stereotypes may unconsciously bring users together on online platforms to form communities. Hence, having linguistic information of a community, such as the topics users in that community interact with and the stance of users towards different pieces of news, can help capture the prevailing stereotypes that form the motivation behind abusive comments from such users.

Demographic characteristics. Previous research has demonstrated that some demographic settings are inherently more abusive than others. For example, a study by Stephens et al. (2013) mapped the locations of hateful tweets across the United States to uncover the regions where people use hate speech the most. They observed that areas with low diversity use more derogatory slurs against racial and sexual minorities. A separate line of work by Savicki et al. (1996) concluded that male-only discussion groups on the Internet use more coarse and abusive language than female-only groups. These works indicate that demographic settings can be predictive of the (abusive) nature of comments originating from within them. User and community information constitutes a direct and simple way of capturing the demographic setting of a comment.

3 Modeling the user and community

In this section, we first recap the datasets in the domain of abusive language detection that contain user or community information alongside comments. We then go on to discuss the methods that have been applied to them.

3.1 Datasets

Twitter has been the most common online platform from which researchers have sourced datasets with user and community information. Galán-García et al. (2016) constructed a dataset of $1,900$ tweets from 19 different twitter accounts with time of publication, language, and geo-position for each tweet taken from the profile of the user who created it. Waseem and Hovy (2016) released a list of $16,907$ tweet IDs along with their corresponding annotations, labeling each tweet as *racist*, *sexist* or *neither*. For each tweet, the dataset contains the gender of the user who created it along with their geo-location. Since Twitter APIs allow researchers to access information about a user given a tweet ID, the dataset of Waseem and Hovy (2016) was expanded by Mishra et al. (2018a) to include the follower-following information amongst users who created the tweets contained in the dataset. Ribeiro et al. (2018) collected a dataset of $100,386$ Twitter users along with up to 200 tweets for each of them. They created a graph of the users based on retweet relationship amongst them and annotated $4,972$ users as hateful or benign based on their tweets. Founta el al. (2018a) released a dataset of $80k$ tweet IDs with labels as *normal*, *spam*, *hateful*, and *abusive*. Augmenting this dataset, Tredici et al. (2019) created a graph of users whose tweets are included based on retweet relationships amongst the them. Similarly, Unsvåg and Gambäck (2018) augmented the datasets of Fortuna (2017) and Ross et al. (2016) which respective contain $5,668$ Portuguese tweets and $13,766$ German tweets by using Twitter APIs to get user information such as gender, number of followers, number of status updates, etc. Deviating from Twitter, Pavlopoulos et al. (2017b) released a dataset of $1.45M$ abusive and benign comments in Greek sourced from the news portal *Gazzetta*. For each comment, the dataset also contains the ID of the user who created the comment.

3.2 Methods

Existing methods for abusive language detection that leverage information of the user who posted the comment or their community can be categorized as social feature engineering based, user embedding based, and social graph based approaches.

3.2.1 Social feature engineering based

These methods directly incorporate hand-engineered features and personal traits of users or their communities in order to model the likelihood of abusive language in the users' comments, a process known as *profiling* (Zhang et al., 2018). Dadvar et al. (2013) included the age of users alongside other traditional lexicon-based features to detect cyber-bullying, while Galán-García et al. (2016) utilized the time of publication, geo-position, and language in the profile of Twitter users. Waseem and Hovy (2016) exploited gender of Twitter users on top of character n-gram counts to improve detection of sexism and racism in a dataset comprising racist, sexist and benign tweets – they noted that the F_1 increased slightly from 73.89% to 73.93% when the gender feature was included. Using the same setup, Unsvåg and Gambäck (2018) showed that the inclusion of social community (i.e., number of followers and friends) and activity (i.e., number of status updates and favorites) features of users alongside their gender further enhanced performance by 3 F_1 points over the n-gram baseline.

3.2.2 User embeddings based

These methods utilize neural networks to generate representations, called *profiles*, for users that capture their linguistic behavior based on the comments they created. Pavlopoulos et al. (2017b) worked with their dataset of abusive and benign comments in Greek. They divided the users whose comments are in the dataset into four *types* based on the proportion of abusive comments: *red* users (e.g., if $>$ 10 comments and $\geq$ 66% abusive comments), *yellow* users (with $>$ 10 comments and 33% $-$ 66% abusive comments), *green* users (with $>$ 10 comments and $\leq$ 33% abusive comments), and *unknown* users (users with $\leq$ 10 comments). They then assigned unique randomly-initialized embeddings to users and added them as additional input alongside representations of comments obtained from the GRU model of Pavlopoulos et al. (2017a). This increased the AUROC from 79.24% to 80.71%. Qian et al. (2018) used LSTMs to model the inter and intra-user relationships in the dataset by Waseem and Hovy (2016) with sexist and racist tweets combined into one category. They first applied a bi-LSTM to users' recent tweets in order to generate intra-user representations that capture the history of their content. To improve robustness against the noise present in tweets, they then utilized locality sensitive hashing to form sets of semantically similar tweets. They trained a policy network to select tweets from these sets that a bi-LSTM could use to generate inter-user representations. When these inter and intra-user representations were utilized alongside representations of tweets from a bi-LSTM baseline, the F_1 score increased from 70.3% to 77.4%.

3.2.3 Social graph based

These methods leverage the social relations (e.g., friendship) that exist amongst users in a social network. Mishra et al. (2018a) constructed a social graph of all the users whose tweets are in the dataset of Waseem and Hovy (Waseem and Hovy, 2016). Nodes were the users and edges the follower–following relationship amongst them on *Twitter*. The researchers applied *node2vec* (Grover and Leskovec, 2016) to this graph to generate representations for users, i.e., *profiles*, which capture information about their social connections. The addition of these profiles on top of linguistic representations of tweets yielded significant gains whereby the F_1 scores on the racism and sexism classes increased from 72.28% and 72.09% to 75.09% and 82.75% respectively. The gains were attributed to the fact that the profiles captured not only information about respective communities of users but also enabled modeling of the topical contexts amongst the connected users. Mishra et al. (2019) further expanded on this work by adding tweet nodes to the social graph of Mishra et al. (2018a) alongside user nodes. They connected every tweet node to the corresponding user who posted the tweet. They then used a graph convolutional network (Kipf and Welling, 2017) to create profiles of users that now captured their linguistic behavior too. When they used these profiles together with the linguistic representations of tweets, F_1 scores on the racism and sexism classes further improved to 79.49% and 84.44% respectively. Ribeiro et al. (2018) also applied graph neural networks, GraphSage (Hamilton et al., 2017), to their social graph of approximately $100k$ Twitter users to generate profiles that they used to classify the users as hateful or normal. They noted that their social graph based method outperformed traditional gradient-boosted decision tree classifiers by 15 F_1 points on the same task. Tredici et al. (2019) constructed a graph of users whose tweets are in the hate-speech dataset of Founta et al. (2018b). Nodes were uses and edges between them signified that one user retweeted the other. They used Graph Attention Networks (Veličković et al., 2018) to generate representations of users from this graph, which when used alongside linguistic representations, provided a gain of 5 F_1 points. Cecillon et al. (2021) worked with a social graph of users from a French gaming website where weighted edges represented the intensity of communication between the users. Then for each comment to be classified, the researchers extracted the ego-graph of its author and created a feature vector for the comment from the ego-graph using *node2vec* along with measures like degree centrality. An SVM trained with these graph-based feature vectors reached 89 F_1 points as opposed to 81 F_1 points when trained with content features.

4 Analysis of the methods

We now analyze the methods described above to understand the gains that user or community information provides. Based on this analysis, in the next sections, we explore the ethical considerations of incorporating user and community information and how it can support explainability.

Across the three categories of methods, we note that the general setup is to create representations, called *profiles*, for users or communities and utilize them alongside linguistic features. In social feature engineering based methods, these profiles are manually constructed vectors of features that capture the relevant traits, such as age in the case of cyber-bullying and gender in the case of sexism. In user embeddings and social graph based methods, the profiles are instead generated by neural network architectures to capture the linguistic behavior or community traits of users. That said, across all three categories, the profiles essentially provide a wider context to the comment being classified for abuse. For example, having the gender of the user who produces a comment such as "*Had an accident, women can't drive it seems!*" can help to classify the comment as sexist or not by differentiating benign self-deprecating humor from intent to degrade. The context that the profiles encode increases as we go from social feature engineering based methods to user embeddings based methods and further to social graph based methods. This is also evident from the magnitude of gains that the profiles provide on top of linguistic features. For example, the gender feature only increases the F_1 from 73.89% to 73.93% over character n-gram counts on the dataset by Waseem and Hovy (2016), while the social graph based method of Mishra et al. (2019) increases the F_1 to above 80%. The example aside, it makes intuitive sense that profiles from social graph based methods encode the most amount of context, since these profiles are able to capture the various phenomena that occur in social networks, the most prominent ones of which are:

- *Homophily*, i.e., the tendency of users in a social space forge ties with others who are similar to them in socially significant ways (McPherson et al., 2001).
- *Coordinated behavior* or *brigading*, i.e., when users with similar beliefs act in a coordinated manner in a social space towards some common objective (Parent et al., 2019).

In fact, homophily is so prominent, Mishra et al. (2019) noted in their work that the profiles they generated from the social graph of users and tweets could encode patters of similar linguistic practices amongst connected users in the Waseem and Hovy (2016) dataset, hence allowing for comments with implicit and generalized sexism or racism to be better detected. Moreover, homophily has direct associations with all the four aspects of context that we described in section 2, i.e., similar sociolinguistic norms and shared language markers facilitate homophilic ties in social networks (Kovacs and Kleinbaum, 2020), as do shared beliefs, stereotypes, and demographic traits (Mishra et al., 2018a). Therefore, capturing homophily allows for all the four aspects to be directly captured together.

We note that just exploiting simplistic and limited inductive biases that are easy to extract, like gender of the user, can render methods prone to making faulty generalizations because of overfitting to patterns in the training data. This is also evident from the observations that Mishra et al. (2019) made in their work. They noted that the profiles they generated from the social graph consisting of user and tweet nodes improved F_1 scores over the profiles Mishra et al. (2018a) generated from the social graph just consisting of users, with the gains mainly coming from increase in precision.

It is because solely relying on *network homophily* as the inductive bias for generating profiles caused the method of Mishra et al. (2018a) to make some faulty generalizations. Such observations have also been made by other works, a prominent one of which is the work of Bamman et al. (2014) who explored the relationships amongst gender, language, and social network connections. The researchers noted that even though there may exist many linguistic clusters that exhibit strong orientations to one gender, yet the characteristics of any particular cluster do not necessarily align with population-level statistics for that gender. Furthermore, they observed that there are individuals whose linguistic practices differ from population-level trends for their gender and that gender homophily does not capture their linguistic practices.

5 Ethical considerations

While researchers have started incorporating user and community information into detection of abusive language, there has been no discussion of the ethical guidelines for doing so. Therefore, taking a stand on the issue, we lay out five ethical considerations in the design and implementation of methods that incorporate user or community information:

Personal vs. population-level trends. It is important to perform appropriate generalizations from personal traits to population-level behavioral trends. Methods should avoid relying on simple inductive biases such as personal traits of users, e.g., gender, race, etc., as this can easily lead to scenarios of faulty generalizations where comments from a particular gender or race are always labeled abusive/benign. Moreover, relying solely on personal traits of users also comes with the risk that such information may not always be present or may not be accurate even when present (Drouin et al., 2016). On the other hand, more complex inductive biases learned from data, as in the case of social graph based methods, provide a safer and more reliable generalization from personal behaviors of users or communities to population level trends.

Bias in datasets. An obvious pitfall in working with methods that incorporate user and community information is having datasets where comments come from users belonging to some limited demographics only. We refer to this as *demographic bias*. Datasets with demographic bias will cause the methods to overfit to linguistic practices and dialects of users and communities belonging to specific demographics (Sap et al., 2020), hence diminishing the power of the methods to generalize. In fact, this bias is not only a problem for methods we discussed, but for any NLP method in general. When it comes to methods that incorporate user or community information specifically, there are two other biases that must be kept in mind when constructing datasets; we refer to them as *comment distribution bias* and *label distribution bias*. Comment distribution bias occurs when the majority of comments in the dataset come from a small number of unique users. Such datasets allow the methods to simply overfit to the linguistic or social behaviors and community roles of specific users (Wiegand et al., 2019). Label distribution bias occurs when only the abusive comments of a user are included in the dataset. Abuse is a relatively infrequent phenomenon, even at an individual level (Waseem and Hovy, 2016; Wulczyn et al., 2017). Only getting abusive comments of a user can make the methods simply associate the identity of the user to abusiveness when including user information. Moreover, datasets with this bias can also make phenomena like homophily appear overly effective in the detection of abuse by sampling only abusive comments from users who are close in the social network.

Observability. The observability aspect needs to be accounted for, i.e., does a method allow for the profiling knowledge it has learned about users and communities to be directly or indirectly observed in its workings, e.g., if it has segregated users into categories observable by others. If yes, that can be used as a basis for systematic oppression of certain users or communities by other users and communities. A prime example of this is when users report benign comments that they do not agree with as abusive since they have noted that the detection method is more likely to adjudicate the comments abusive simply because they come from a particular community or a particular user.

Privacy. As we discussed in the previous section, profiles created by the methods may carry a lot of information about the personal traits of users, their linguistic practices, etc. Furthermore, the information carried increases in specificity as we go from social feature engineering based methods to social graph based methods. An important ethical consideration that then arises is whether the profiles or the models learned by the methods be made available publicly. Doing so may allow for users and communities to be uniquely identified and for their sociolinguistic behaviors, community roles, or personal and population-level beliefs to be exposed.

Purpose. The purpose of leveraging user and community information should be made clear upfront. Methods that leverage user and community information to enhance the detection of abusive language in comments should be preferred over those that leverage the information to classify users or communities themselves as abusive. This is because the latter can lead to unwarranted penalties, e.g., a platform may prohibit a user from engaging even in restorative conversations simply because of their past abusive behavior.

6 Explainable abusive language detection

Explainability is an important concept within abusive language detection. Jurgens et al. (2019) noted in their work that explainable ML techniques can promote *restorative* and *procedural* justice by surfacing the norms that have been violated and clarifying how they have been violated. That said, there has been limited discussion of the issue within the domain of abusive language detection. In this section, we first formalize the properties that an explainable detection method should aim to exhibit in order to thoroughly substantiate its decisions. We then describe how user and community information play an important role in the realization of each of the properties. Finally, we discuss what it means to operationalize explainability within abusive language detection in an effective manner.

6.1 Properties of an explainable method

In drawing up the properties that an explainable method for abusive language detection should aim to exhibit, we take into account the taxonomy of abuse we discussed in the introduction, i.e., directed vs. generalized and implicit vs. explicit:

- *Provide evidence for intent* of abuse (or the lack of it), hence convincingly segregating abuse from other phenomena such as sarcasm and humor.
- *Point out the abusive phrases* within a comment (or the absence thereof), be they explicit (e.g., expletives or slurs) or implicit (e.g., dehumanizing comparisons).
- *Identify the target(s)* of abuse (or the absence thereof), be it an individual (i.e., directed abuse) or a group (i.e., generalized abuse).
- *Elucidate stereotypes(s)* underlying the abuse (or the absence thereof), be they explicit or be they in the form of implicit associations.

User and community information has a crucial role to play in the effective realization of each of the four properties. For the first property, as illustrated earlier in the paper, information of the user who created the comment can serve as evidence for whether the comment intends to be degrading to others or just self-deprecating humor. For the second property, let us consider a comment like "*You're a pig!*"; if directed at people belonging to certain religions, it may constitute an implicit racial slur, but otherwise, may simply be viewed as a remark on cleanliness. So, the information of the user or community being targeted can explain whether a phrase is abusive or not. The methods we analyzed in section 4 do not model the information of the target user or community, which is a valuable direction for future research. For the third property, we note that social graph based methods are inherently suited to provide a convenient setup for identification of the user or community being targeted by an abusive comment, specially in scenarios where the social graph is enriched with information like the topics being discussed amongst groups of connected users. For the fourth property, user and community information again offers a direct way to elucidate explicit or implicit stereotypes, e.g., by exposing the associations being made by a community between certain qualities and the targets of their abuse.

6.2 Operationalizing explainability

Having formalized the properties that an explainable detection method should aim to exhibit, we now address the question of how explainability can be effectively be operationalized within abusive language detection in view of these properties. We approach this discussion from three different perspectives, that of the designers of the detection method, that of the user creating comments, and that of the larger communities. By breaking the discussion down in this manner, we explore the different choices that exist for operationalization and the purposes they can serve.

Designers of the method. For the designers of the detection method, explainability can serve as a principled mechanism for understanding and reasoning about the behavior of their method, which is important for multiple reasons. Firstly, if the detection method exhibits all the four properties of explainability, then the designers can easily gain insights into the factors that contributed to the decision made by the method given a comment. This can allow the designers to recognize when the method may be overly relying on a specific factor, e.g., the demographic traits. In the case of social feature engineering and user embeddings based methods, operationalization of explainability via feature attribution such as *LIME* (Ribeiro et al., 2016) and *Integrated Gradients* (Sundararajan et al., 2017) can be effective in offering such insights. For social graph based methods that employ graph neural networks, attribution techniques like *GNNExplainer* (Ying et al., 2019) can be used instead. The second reason why explainability is important for the designers is because it can allow them to optimize the method by removing inputs that do not contribute significantly. Here again, explainability via feature attribution can be effective. Lastly, explainability is also important for the designers to understand how their method would perform in cases where a user may try obfuscate abusive language (Nobata et al., 2016). Counterfactual explanations can constitute an effective operationalization for the designers to identify the parts of their method that are most vulnerable to obfuscations.

Users. Besides being a mechanism for designers to interpret their methods, an effective operationalization of explainability should also serve as a means for users to receive explanations for the decisions made by a detection method. Jurgens et al. (2019) argue in their work that an online platform can build legitimacy and transparency by offering justifications to users when their comments are deemed abusive by the detection method of the platform, which can in turn lead to increase in compliance with the norms of the platform. That said, unlike in the case of designers of the method, offering feature attribution based explanations that simply highlight parts of a user's comment may not be effective at making the user agree with the decision of the detection method (Carton et al., 2020). Alternatively, providing a meaningful counterfactual paraphrase that is non-abusive is not only difficult (Laugel et al., 2019), but can also be seen as *paternalism* on the part of the platform (Barocas et al., 2020), i.e., that the platform is trying to tell the user what to say or how to present their opinions. On the other hand, *principal-reason explanations* (Barocas et al., 2020), whereby the detection method selects the reason(s) for its decision from a curated list, can constitute an effective operationalization. Such a list can be prepared for each of the four properties of explainability, e.g., by selecting the relevant norms from the *terms of service* of the platform, hence allowing for a principal reason to be offered per property or a combination thereof. When coupled with feature attribution, this approach to operationalization can clearly indicate to the user the norm(s) that their comment violates and, where possible, highlight parts that contribute to the violation(s). For example, given a comment like "*You f***, why do you have to support that team??*", the detection method can highlight the first part based on feature attribution and select the norm forbidding the use of expletives directed at others.

Communities. There can be scenarios where whole communities of users on a platform may be indulging in abusive behavior, e.g., by widely circulating an abusive view against a demographic group based on shared beliefs, common stereotypes or other homophilic ties. In such cases, just taking down specific instances of abusive language and providing justifications individually to the respective users may not prove effective. Users may continue to promote the abusive view, defying the norms of the platform in the process and ignoring the justifications given to them. The reason for this comes from *social influence theory* which says that a user's behavior is affected by three broad varieties of social influence (Kelman, 1958), i.e., *compliance*, *identification*, and *internalization*. Compliance occurs when the user behaves a certain way so as to appear in congruence with opinions of others who matter to them; identification occurs when the user adopts behaviors in order to associate with others they admire; and internalization is when the user adopts the values and beliefs of others. The influences occur because of two needs of the user, the need to be liked (*normative*) and the need to be right (*informational*). In order to fulfill the latter, people may accept the three varieties of influence when there is lack of information, a concept known as *social proof* (Cialdini, 2007). Consequently, explainability has a bigger role to play here than simply being a tool that provides interpretability to designers or offers justifications to users. Operationalizing explainability in a manner that spreads awareness about existing stereotypes and fills the information gap can be very effective (Miller, 2018; Sap et al., 2020). One way to achieve this is by having generative explanations in conjunction with information retrieval techniques that fulfill the property of elucidating stereotypes in a human-understandable way (Gilpin et al., 2018) while offering references to reliable sources on the stereotypes. In fact, such an operationalization that elucidates stereotypes or frames of bias (Sap et al., 2020) in abusive comments at a community level, while providing information to debunk the stereotypes themselves, can offer validation to the victims of abuse by communities, e.g., minority groups, and help them feel safer on the platform.

7 Conclusions

Abuse on the Internet stands as a significant challenge before the society. Its nature and characteristics constantly evolve, making it a complex phenomenon to study and model. In this paper, we explored the ways in which users and communities play a role in the detection of abusive language. We investigated the methods that leverage user or community information to uncover how they work and the knowledge they capture. We then explored the ethical challenges of incorporating user and community information, laying out considerations to guide future research. Finally, we moved to the topic of explainability in abusive language detection, proposing properties that an explainable detection method should aim to exhibit. We describe how user and community information can facilitate the realization of these properties and discussed the effective operationalization of explainability in view of the properties.

References

I. Awan. 2014. Islamophobia and twitter: A typology of online hate against muslims on social media. *Policy & Internet*, 6:133–150.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *FAT* '20*, page 80–89, New York, NY, USA. Association for Computing Machinery.

Anat Ben-David and Ariadna Matamoros Fernández. 2016. Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in spain. *International Journal of Communication*, 10(0).

Nick Byrd. 2019. What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese*, pages 1–29.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-based explanations don't help people detect misclassifications of online toxicity. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):95–106.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32.

Robert B Cialdini. 2007. *Influence: The psychology of persuasion*, volume 55. Collins New York.

Noé Cécillon, Vincent Labatut, Richard Dufour, and Georges Linarès. 2021. Graph embeddings for abusive language detection. *SN Computer Science*, 2(1).

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 693–696.

Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4707–4717, Hong Kong, China. Association for Computational Linguistics.

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*, WWW '15 Companion, pages 29–30. Association for Computing Machinery.

Michelle Drouin, Daniel Miller, Shaun M.J. Wehle, and Elisa Hernandez. 2016. Why do people lie online? because everyone lies on the internet. *Comput. Hum. Behav.*, 64(C):134–142.

Maeve Duggan. 2017. Online harassment 2017.

Alexandra D'Arcy and Taylor Marie Young. 2012. Ethics and social media: Implications for sociolinguistics in the networked public1. *Journal of Sociolinguistics*, 16(4):532–546.

Paula Fortuna. 2017. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018a. Large scale crowdsourcing and characterization of twitter abusive behavior. In *International AAAI Conference on Web and Social Media*.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018b. Large scale crowdsourcing and characterization of twitter abusive behavior. In *International AAAI Conference on Web and Social Media*.

Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2016. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034.

P. Hinton. 2017. Implicit stereotypes and the predictive brain: cognition and culture in "biased" person perception. *Palgrave Communications*, 3:1–9.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *International AAAI Conference on Web and Social Media*.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

H.C. Kelman. 1958. Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution*, 2(1):51–60.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17.

Balazs Kovacs and Adam M. Kleinbaum. 2020. Language-style similarity and social networks. *Psychological Science*, 31(2):202–213. PMID: 31877069.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The dangers of post-hoc interpretability: Unjustified counterfactual explanations.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303. Association for Computational Linguistics.

Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018a. Author profiling for abuse detection. In *Proceedings of COLING 2018*, pages 1088–1098. Association for Computational Linguistics.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive language detection with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Emily R. Munro. 2011. The protection of children online: a brief scoping review to identify vulnerable groups.

Dong Nguyen and Carolyn P. Rosé. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Language in Social Media (LSM)*, LSM '11, page 76–85, USA. Association for Computational Linguistics.

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*, WWW '16, pages 145–153. The International World Wide Web Conference Committee.

Mike C. Parent, Teresa Gobble, and Aaron Rochlen. 2019. Social media behavior, toxic masculinity, and depression. *Psychology of Men and Masculinity*, 20:277–287.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the 1st Workshop on Abusive Language Online*, pages 41–45. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deeper attention to abusive user content moderation. In *Proceedings of EMNLP 2017*, pages 1125–1135. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. 2017b. Improved abusive comment moderation with user embeddings. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 51–55. Association for Computational Linguistics.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the NAACL: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123. Association for Computational Linguistics.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, AI'10, pages 16–27.

Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *International AAAI Conference on Web and Social Media*.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Ben 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 NLP for Computer-Mediated Communication*, pages 6–9.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *International AAAI Conference on Web and Social Media*.

Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. 2019. Attending the emotions to detect online abusive language. *arXiv preprint arXiv:1909.03100*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Victor Savicki, Dawn Lingenfelter, and Merle Kelley. 1996. Gender Language Style and Group Composition in Internet Discussion Groups. *Journal of Computer-Mediated Communication*, 2(3). JCMC232.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, pages 1–10. Association for Computational Linguistics.

Sara Owsley Sood, Judd Antin, and Elizabeth F Churchill. 2012. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume 12, page 06.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of the 14th AAAI and 9th IAAI*, AAAI'97/IAAI'97, pages 1058–1065. AAAI Press.

Monica Stephens. 2013. Mapping the geography of hate. [Online; accessed 26 January 2021].

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the 2nd Workshop on Language in Social Media*, LSM '12, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the 1st Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL SRW*, pages 88–93. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056. Association for Computational Linguistics.

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*, WWW '17, pages 1391–1399. The International World Wide Web Conference Committee.

Dawei Yin, Brian D. Davison, Zhenzhen Xue, Liangjie Hong, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. In *Processings of the Content Analysis in the WEB 2.0*.

Rex Ying, Dylan Bourgeois, Jiaxuan You, M. Zitnik, and J. Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32:9240–9251.

Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan. 2018. *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part II*, volume 11109. Springer.

Matthew Zook. 2012. Mapping racist tweets in response to president obama's re-election. [Online; accessed 26 January 2021].