Online Hate Detection Systems: Challenges and Action Points for Developers, Data Scientists, and Researchers

Joni Salminen
Qatar Computing Research Institute
Doha, Qatar
jsalminen@hbku.edu.qa

Maria Jose Linarez
Universidad Centroccidental Lisandro Alvarado
Barquisimeto, Venezuela
linarezcastillo86@gmail.com

Soon-gyo Jung
Qatar Computing Research Institute
Doha, Qatar
sjung@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute
Doha, Qatar
bjansen@hbku.edu.qa

***Abstract*—Automated online hate detection has garnered interest from various stakeholders to make online platforms safer. Despite this interest, there remain a plethora of unresolved issues that hinder advancement. We review fourteen state-of-the-art articles discussing these challenges, and present a meta-synthesis. Six themes are identified: (1) Dataset selection, (2) Detection of False Positives and Negatives, (3) Semantic Context of Hate Messages, (4) Privacy and Anonymity, (5) Ethical Considerations, and (6) Minimizing Bias. For each theme, we provide a set of action points to support researchers, data scientists, and developers to improve hate detection systems.**

Keywords—Online hate, social media, hate detection systems

I. INTRODUCTION

Online communication is characterized by being open, public, and often anonymous. The negative side-effect of openness is that some users have taken advantage of the condition of mobility and pseudo-anonymity on social media to express messages of hate and intolerance [1]–[4]. Although unfortunate, it appears to be a de facto condition of the medium. Users may post abusive messages that can cause controversies and moral damage to a person or groups of people [5], [6], who could consequently feel self-conscious when expressing themselves, for fear of being harassed or attacked, or stop participating in online conversations [7] due to fear of being targeted. Central concepts in this line of work are online toxicity and hate:

- **Online toxicity** is defined as harmful (intentionally hurtful) communication via public or private messaging that discourages users to participate in public discussion and/or encourages them to leave the platform or discussion forum altogether [4].
- **Online hate** is defined as harmful communication that is targeting an individual or a group [8]. When the target is an individual, the term 'cyberbullying' applies; when the target is a group, hate speech is used. The target of hate can also be an institution, such as the media or police, or humanity itself [3].

Overall, hate speech and other forms of online hate (cyberbullying, toxicity, abusive language [9]) negatively affect online communication and equal participation in social media. For this reason, the development of algorithms for detecting online hate has been promoted by social media platforms and other stakeholders supporting inclusive and constructive communication among all online users.

Classification algorithms from natural language processing (NLP) are typically used for the detection of online hate [10]. In this task, algorithms face various challenges. Some of these challenges are related to the design of algorithms in general, such as the correct selection of the dataset and the detection of false positives [11]. Other challenges involve interpretation, such as evaluating and disambiguating the semantic context of hateful messages and mitigating bias [12]. The challenges also involve ethical aspects, such as privacy and anonymity, ethics of the data collection, and individuals' freedom of expression [13].

While a number of studies investigate these challenges, there is a lack of understanding of what the challenges mean for the development of algorithm-based hate detection systems. Therefore, a clear picture of *what should be done* is missing from the body of knowledge. We address this gap by providing a meta-synthesis of the previous literature and considering the perspectives of three stakeholder groups: (1) *researchers* that use the detection models to study online hate as a phenomenon, (2) *data scientists* that train hate classifiers, and (3) *system developers* that implement the hate classifiers and algorithms in real systems, applications, and user interfaces (UIs). Researchers and data scientists are mainly focused on *creating* hate detection models and algorithms, whereas the latter focuses on *implementing* these models and algorithms in real systems and applications. In this research, we suggest specific action points (AP) for each stakeholder group based on generally known challenges of online hate detection.

II. METHODOLOGY

We used relevant keywords to search Google Scholar for articles to review, including 'online hate/toxicity/abuse', 'challenges', 'errors', 'problems', "literature review", '+detection'. We combined these to form search phrases, such as ["online hate" +challenges +detection], ["online toxicity" +problems +detection], and so on. Altogether, twelve searches were conducted. As there were hundreds or thousands of results for each search phrase (we did not record exact numbers, as these are prone to frequent change with new research published continuously), the results were too numerous to manually screen. Therefore, a heuristic rule was applied to screen every result in the first five search result pages in Google Scholar. The screening was done by reading the titles and abstracts, using these criteria:

- **Relevance**—title and abstract mention a challenge or challenges concerning online hate detection
- **Recency**—we wanted to emphasize state-of-the-art challenges, so articles prior to 2017 were excluded

Fourteen articles passed the screening and were thus included. The selected 14 articles (see Table 1) were

2021 8th International Conference on Behavioral and Social Computing (BESC) | 978-1-6654-0023-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/BESC53957.2021.9635436

published between 2017 and 2020; one (7.1%) article in 2017, three (21.4%%) in 2018, six (42.9%%) in 2019, and four (28.6%%) in 2020. Six (42.9%) were review articles, seven (50.0%) empirical, and one (7.1%) conceptual.

TABLE I. THE INCLUDED ARTICLES, SORTED BY PUBLICATION YEAR. TYPE: C = CONCEPTUAL, E = EMPIRICAL, R = REVIEW.

ID	Article title	Year	Type
01	Tracking Hate in Social Media: Evaluation, Challenges and Approaches [14]	2020	R
02	Directions in Abusive Language Training Data: Garbage In, Garbage Out [11]	2020	R
03	Evaluating Semantic Feature Representations to Efficiently Detect Hate Intent on Social Media [15]	2020	E
04	Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation [16]	2020	E
05	Tackling Online Abuse: A Survey of Automated Abuse Detection Methods [17]	2019	R
06	Hate speech detection: Challenges and solutions [18]	2019	R
07	Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate [19]	2019	R
08	Trends in the Regulation of Hate Speech and Fake News: A Threat to Free Speech? [20]	2019	C
09	The Thin Line Between Hate and Profanity [21]	2019	E
10	Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter [22]	2019	E
11	Hate Speech Detection on Twitter: Feature Engineering vs. Feature Selection [23]	2018	E
12	Challenges for Toxic Comment Classification: An In-Depth Error Analysis [24]	2018	E
13	A survey on automatic detection of hate speech in text [9]	2018	R
14	The bag of communities: Identifying abusive behavior online with preexisting internet data [25]	2017	E

III. CURRENT CHALLENGES IN ONLINE HATE DETECTION

We downloaded the full-text versions of the articles mentioned in Table 1, which we then fully read to synthesize the main challenges currently facing online hate detection. Based on a thorough reading, we were able to classify the different challenges into six themes as shown in Table 2.

A. Dataset Selection

The correct selection of the training dataset for an algorithm is essential to obtain useful results. In this sense, Vidgen and Derczynski [11] state that *the detection of online hate will be effective to the extent that the correct data are used*. However, curating and selecting datasets is not necessarily easy, since the data must satisfy several conditions, among which are that data must not be biased, be considerably large and have a theoretical foundation. In addition, there is a large number of publicly available datasets to choose from [11].

Vidgen and Derczynski [11], with the initiative of helping to improve the selection of the data set for algorithm training, have summarized a series of recommendations that help design a dataset that can be adjusted to the desired requirements. Among the recommendations that the authors consider useful is the *documentation* of the dataset created, in addition to being clear about the purpose for which the data will be used. They consider that this is an area that deserves permanent development and research, providing tools to create accessible and truly useful datasets. This is epitomized in the catchphrase "Garbage in, garbage out".

The choice of the dataset to train data is a general challenge for the predictive accuracy of machine learning algorithms. Particularly, in the area of hate speech detection, a dataset that does not represent all types of hateful language or all targets of hate can lead to biased or inaccurate results. In this way, it is presented as pertinent to develop adequate techniques to improve the selection of the dataset [14].

Regarding online hate datasets, many researchers currently agree that a specific challenge is to *include the greatest diversity of languages possible*, apart from making adjustments that improve the already existing algorithms, which are mostly in English. In this regard, Arango et al. [16] conducted a study that handled the data in both English and Spanish, this being an important initiative for future.

It is also important to highlight the preprocessing of the training data. When data contains a lot of noise, algorithms cannot be properly trained. For example, Madukwe and Gao [21] point out that in the case of Twitter, the data contains a large number of characters that do not add any value to the analysis. Data cleaning and preprocessing can help in having a more useful dataset for model training.

In addition to data cleaning and preprocessing, feature extraction or engineering is crucial – referring to how text is transformed into numerical format for the algorithms [23].

According to Sahlren et al. [26], the task of online hate detection is defined by finding efficient feature representations, which in turn is driven by the general progress of NLP. At the same time, this implies that datasets and feature representation need to be frequently updated to reflect changes in linguistic expressions of hate [9] as well as novel feature representations. As an example, Deep Bidirectional Transformers (BERT) [27] did not exist three years ago, but today they are part of virtually all the state-of-the-art hate detection models [10].

Based on these concerns, we propose the following **Action Points (APs)**:

- **AP01:** Online hate datasets are many and the state of their quality is often unclear. Third-party studies are required to verify the quality of the datasets. [*researchers, data scientists*]
- **AP02:** Develop multi-lingual hate classifiers. [*researchers, data scientists*]
- **AP03:** Acquire new samples periodically to refresh the training sets for model development. [*data scientists*]
- **AP04:** Continuously monitor and test NLP developments for more effective feature representation. [*researchers, data scientists*]

TABLE II. CHALLENGES OF ONLINE HATE DETECTION RAISED BY DIFFERENT STUDIES. 'X' INDICATES THE CHALLENGE IS MENTIONED IN THE PAPER.

Paper ID	Dataset selection	Detection of false positives and negatives	Semantic context	Privacy and anonymity	Minimizing bias	Implementation
01	X			X	X	X
02	X			X	X	
03	X		X			
04						X
05	X		X		X	
06	X	X				
07	X		X			
08	X					
09		X	X			
10						
11		X		X		
12		X				
13						
14			X	X		X

B. Detection of False Positives and Negatives

Van Aken et al. [24] used a set of last-generation classifiers, considering that the optimal handling of false positives and negatives also depends on how well the labels are defined, the training of the models, and the number of elements of the classes of the training datasets. In general, when working with small classes, erroneous results are obtained. Robinson et al. [23] and Zhang and Luo [22] have investigated data from Twitter, which is perhaps the most widely analyzed platform of online hate [9]. The researchers postulate that efforts should be made to improve the techniques used to detect offensive messages, and that the *techniques used for early detection of false positives and negatives would make hate detection systems more robust*.

In their effort to minimize the classification error, Madukwe and Gao [21] observed that many messages were misclassified when lexical features alone were used. In some cases where specific words were defined as offensive, the model wrongly classified the message that contained it as hateful. Similarly, some forms of hate can be subtle or obfuscated [18], in that no common "hateful" words are used but the meaning, to a human, is still hateful.

Finally, social media users can actively try and prevent detection of hateful terms, e.g., by writing "m-o-r-o-n" instead of moron. Such adversarial examples can fool an otherwise robust classifier [28]. Therefore, it can be concluded that *a message that contains certain words whose lexical meaning is hate does not necessarily represent a hateful message*. On similar lines, *a message that does not contain certain words whose lexical meaning is hate can still represent a hateful message*. Thus, the tendency of algorithms to focus on specific words or expressions leads to the presence of false positives and negatives. This is inevitably the case, we might add, as none of the known classifiers have ever reached a perfect classification score. Given that the lexical meaning of certain words is used in the training of algorithms, the algorithms must be better calibrated to determine this type of false positive.

Based on these concerns, we propose the following **APs:**

- **AP05:** Apply early detection of false positives and negatives. [*researchers, data scientists*]
- **AP06:** Conduct "sanity checks" on the results by asking the model to predict non-hateful sentences that contain a typically hateful word. [*researchers, data scientists, developers*]
- **AP07:** Consider ways of dealing with the inevitable false positives and negatives [*developers*]

C. Semantic Context of Hate Messages

When hate detection systems rely on the outputs given by algorithms, it is particularly important to bear in mind the semantic meaning of the sentences being analyzed. Classification algorithms must be trained so that they can detect the meaning of a sentence, not so much by the occurrence of certain words but by the semantic relationship between them. In other words, *classifiers that are able to interpret the semantic context of phrases perform better for hate detection than those that cannot*. This is why methods such as BERT provide a performance boost, as they are designed for capturing semantic relationships [27].

The requirement to understand context goes hand in hand with linguistics (i.e., the analysis of language), so that ambiguity in the use of certain terms can be avoided. This is noted by Van Aken et al. [24] who present an important challenge for detection algorithms: *the lack of training data with highly idiosyncratic or rare vocabulary*. Mishra et al. [17] state that considering the context of individual messages is a key aspect for the effective detection of online abuse (a.k.a., hatred, profanity, toxicity). They found that analyzing the content of the messages without considering the various meanings that an expression can have depending on the context, can miss the detection of abuses that are serious and offensive for a user. For example, figurative language is a very challenging research task in NLP, where hate tone is observed yet lacking effective approaches to be detected.

On the other hand, the researchers affirm that the detection of online abuse is an issue that becomes more complex, since users often use metaphors, sarcasm, comparisons, images (that is, a wide range of expressions that can be used for an offense. Since there are variations of communicating a message, it would be necessary to strengthen the versatility of tools for detecting tacit meanings and apply common sense [11].

Considering that online abuse is a subject in constant flux, Chandrasekharan et al. [25] claim that there are characteristics considered in the models, which become negligible as time passes, that is, lose relevance. It is rare that

researchers revisit their machine learning models after they are published [29], which implies that *using research-based hate detection models in production may not be a viable option in the long run*. Taking these aspects into account, Modha et al. [14] ensure that the context of the messages must be frequently evaluated and the models updated so that they are capable of detecting sarcasm, humor, youth slang, and satire that are sometimes used with offensive purposes.

Senarath and Purohit [15] state that the careful study of the semantic context of messages is essential for the design of effective algorithms for detecting online hate. They also place special emphasis on the need for the subtlety of certain offensive messages to be detected. Sahlgren et al. [26] refer to this as the obfuscated hate problem.

Polysemy refers to words with multiple meanings, which can lead to algorithms misclassifying a message. Although the representation of features is refined by Senarath and Purohit [15], it is clear that the *detection of different meanings in certain polysemic words needs to be improved in the future*, so that algorithms can correctly classify messages that present ambiguity.

Based on these concerns, we propose the following **APs:**

- **AP08:** Consider ways of varying the context of hateful/non-hateful expressions when creating the training data (e.g., find specific examples of obfuscated hate) [*researchers, data scientists*]
- **AP09:** There is a need to systematically investigate how robust a proposed hate classifier is against multiple forms of linguistic nuances, including polysemy, humor, sarcasm, and satire. [*researchers, data scientists*]

D. Privacy and Anonymity

It is relevant to preserve the privacy of user data. Particularly in environments where data related to hate speech is handled, privacy is particularly valuable, given that the possibility of linking the identity of a subject who makes (or is the target of) a negative comment with said comment can lead to the risk of the physical integrity of the individual. Besides, there is the possibility that he will be subjected to public derision or vilification as a result of having made a hateful comment. In this sense, research on this topic must be particularly careful that the privacy of users is preserved. Vidgen and Derczynski [11] and Mishra et al. [17] agree that the privacy of users should not be compromised, to improve the algorithms for the recognition of hate speech.

Special mention deserves the approach of Arango et al. [16], where an antagonistic vision is established between the need to preserve privacy and the scientific value of knowing the source of the message. According to law, the privacy of the users must be guaranteed, but a particular scientific value is also given to knowing if several hate messages come from the same source. If privacy is strictly preserved, it is difficult to determine the distribution of users, which can hamper the accuracy of hate or abuse detection models and algorithms.

Privacy also represents a challenge for detection algorithms, since the platforms where messages are exchanged between users are obliged not to reveal user data, preventing researchers from having access to a greater amount of information that can serve to improve the detection models. This *scarcity of data prevents progress in the improvement of the algorithms and the general advancement of this area of research* [11].

On the other hand, certain platforms allow users to make comments anonymously, with the largest presence of anonymous communications being Facebook, followed by YouTube and Twitter [20]. This issue of anonymity is addressed by Zhang and Luo [22], who show how the anonymity and mobility that are possible in media such as Twitter, have meant an increase in hate speech. Therefore, *anonymity prevents the profiling of users and makes it difficult to detect hate speech in many cases*.

Based on these concerns, we propose the following **APs:**

- **AP10**: Remove personally identifiable information from online hate datasets that are shared publicly. [*researchers*]
- **AP11**: Apply user IDs to retain source consistency without the need for using personally identifiable information such as name or username. [*researchers, data scientists*]
- **AP12**: Protect the privacy of perpetrators and targets alike. [*researchers, data scientists, developers*]

E. Minimizing Bias and Maximizing Generalizability

In the development of systems based on probabilistic algorithms, the challenge of managing bias for the estimators is always present, and in the case of detecting hate, it should be taken as a starting point that online hate is not a binary problem [30], but it includes multiple dimensions, such as racism, sexism, religion, disability, and gender [21].

Bias is defined as the presence of errors in the classification algorithms, that is, making wrong classifications based solely on the data features and giving misleading results. In the context of hate systems, the possible dangers of having a high bias are that algorithms mistakenly associate online abuse with certain people unfairly considering only aspects such as race, religion, gender, or political affiliation.

Arango et al. [16] maintain that it is important to reduce bias; they propose that it is essential to investigate where the data used comes from, to avoid as much as possible working with biased information, in the case of computer-mediated communication, it is important to know if part of the analyzed messages come from the same user or how diverse the data studied is. Working with a data set with high bias is the cause of obtaining very optimistic results of the performance of the algorithms, which represents a problem given that when using these algorithms for other data, their result often does not correspond to reality, that is, there is *an algorithm that detects hate and toxicity for a specific data set and does not work for other data*.

On the other hand, Vidgen and Derczynski [11] state that currently detecting online abuse in real-time, without bias, is a priority for technology companies, political sectors, and researchers in this area of knowledge. Also, they present their ideas on the task of creating datasets for the training of such algorithms, considering that the data is one of the most relevant aspects for controlling bias in abuse detection algorithms. Overall, debiasing hate detection systems is a

socio-technical challenge, requiring human participation in interpreting the results given by the algorithms.

Based on these concerns, we propose the following **APs:**

- **AP13**: Collect and test datasets from multiple social media platforms. [*researchers, data scientists*]
- **AP14**: Develop classifiers that consider multiple targets of hate (i.e., cross-target hate detectors). [*researchers, data scientists*]
- **AP15**: Document choices related to the online hate classifier development and system functionalities based on the classifier to make these choices transparent and open to scrutiny for third parties. [*researchers, data scientists, developers*]

F. Implementation

Online hate speech analysis must address implementation considerations to respect the user's identity. How the data is collected and how the analysis algorithms are processed can potentially represent a risk in the violation of integrity, since the information that is handled is sensitive in many cases.

Collecting information about the profile of users can be benignly motivated if the intention to use it is to better understand the content produced by them. But if the purpose is to take action against the users, then the ethics in data collection would be violated [17]. Vidgen and Derczynski [11] point out that *an ethical way of sharing and using online hate datasets has not been developed so far*. Advances in technology, legislation, and social media platforms' terms of services have not allowed online hate researchers to effectively determine how they can treat sensitive data with a certain ethic to protect the users behind the comments.

An in-depth study of the ethical challenges found for hate detection algorithms is found in [17], where it is shown that the research community is making progress in defining the profiles of users, but without taking into account ethical considerations for it. Although they do not establish a clear answer to how to resolve this ethical issue, they hope that at least the issue will remain open for discussion and be addressed in the future. Ethics must be present throughout the online hate investigation process and at all stages, particularly in the profiling of users. *Anonymity must be preserved at all times, avoiding bias based on users' profile as much as possible, promoting the visibility of profiling so that users are aware of how they were profiled, and finally, be clear about the intention of profiling users*. If these aspects are taken into account, it is possible to advance in an investigation where ethics is present [17].

At the same time, rules and terms of services set by the platforms may decrease researchers' chances of building efficient classifiers. For example, if meta-data or disaggregated data on users is not shared by the platforms, any datasets constructed by researchers will omit these potentially influential variables [11]. Therefore, a common playbook of rules is needed to safeguard both the researchers' ability to effectively study the phenomenon of online hate, while respecting the rights of the users. A partial solution, applied by some companies, is to share their data under non-disclosure agreements (NDA) [31].

Finally, an aspect of implementing the hate detection models is freedom of expression. Santuraki [20] studied the risk that by preventing the spread of toxic and abusive messages that promote hate online, laws are proposed that affect people's free expression. Although each state must guarantee the protection and the dignity of its inhabitants and promote peace within its territory, it must be ensured that the measures adopted do not correspond to the benefits of the government that may seek to silence opinions contrary to them. In this light, online hate regulation falls under the complex territory of information dissemination [32], i.e., setting rules for social media platforms and other information channels while ensuring individuals' freedom of expression.

Based on these concerns, we propose the following **APs:**

- **AP16**: Conduct an in-depth error analysis to carefully analyze any systematic errors given by the online hate classifier. [*researchers, data scientists*]
- **AP17**: Share resources, including online hate datasets, algorithms, and models that enable others to replicate results and build upon them. [*researchers*]
- **AP18**: Measure and display overall error rates of online hate classifiers to end-users such as moderators. [*developers*]
- **AP19**: Provide probabilities and explanations of labels provided by the hate classifier for individual samples. [*developers*]

IV. Research Contribution and Limitations

Our contribution is that while many of the prior reviews mention several online hate detection challenges, none mention their implications to development of hate detection systems. The value of our research comes from synthesizing the challenges from prior work into six themes and offering nineteen action points that different stakeholder groups can pursue to solve or mitigate the challenges when contributing to the development of hate detection systems. Fifteen action points specifically concern researchers, fourteen concern data scientists, and six concern developers (see Table 3). The challenges of online hate detection thus relate to all aspects of systems deploying online hate detection, from the model creation to its implementation (backend, frontend/UI), and integration into human activities (decisions).

Regarding the limitations, the reader should note that the stakeholders in square brackets represent the primary stakeholders for a given action point – this does not mean a given action point would not be relevant for all stakeholders in some sense. From a collaborative perspective, it is possible that a working group, or a taskforce, will carry out different professions in one system, i.e., a researcher can also be a data scientist and design the system. The collaborative aspects of hate detection therefore warrant more research.

TABLE III. ACTION POINTS FOR RESEARCHERS, DATA SCIENTISTS, AND DEVELOPERS

Researchers	Data scientists	Developers
AP01: Online hate datasets are many and the state of their quality is often unclear. Third-party studies are required to verify the quality of the datasets	**AP01:** Online hate datasets are many and the state of their quality is often unclear. Third-party studies are required to verify the quality of the datasets	N/A
AP02: Develop multi-lingual hate classifiers.	**AP02:** Develop multi-lingual hate classifiers.	N/A
	AP03: Acquire new samples periodically to refresh the training sets used for model development.	N/A
AP04: Continuously monitor and test NLP developments for more effective feature representation.	**AP04:** Continuously monitor and test NLP developments for more effective feature representation.	N/A
AP05: Apply early detection of false positives and negatives.	**AP05:** Apply early detection of false positives and negatives.	N/A
AP06: Conduct "sanity checks" on the results by asking the model to predict non-hateful sentences that contain a typically hateful word.	**AP06:** Conduct "sanity checks" on the results by asking the model to predict non-hateful sentences that contain a typically hateful word.	**AP06:** Conduct "sanity checks" on the results by asking the model to predict non-hateful sentences that contain a typically hateful word.
		AP07: Consider ways of dealing with the inevitable false positives and negatives
AP08: Consider ways of deliberately varying the context of hateful/non-hateful expressions when creating the training data	**AP08:** Consider ways of deliberately varying the context of hateful/non-hateful expressions when creating the training data	N/A
AP09: There is a need to systematically investigate how robust a proposed hate classifier is against multiple forms of linguistic nuances, including polysemy, humor, sarcasm, and satire.	**AP09:** There is a need to systematically investigate how robust a proposed hate classifier is against multiple forms of linguistic nuances, including polysemy, humor, sarcasm, and satire.	N/A
AP10: Remove personally identifiable information from online hate datasets that are shared publicly.		N/A
AP11: Apply user IDs to retain source consistency without the need for using personally identifiable information such as name or username	**AP11:** Apply user IDs to retain source consistency without the need for using personally identifiable information such as name or username	N/A
AP12: Protect the privacy of perpetrators and targets alike	**AP12:** Protect the privacy of perpetrators and targets alike	**AP12:** Protect the privacy of perpetrators and targets alike
AP13: Collect and test datasets from multiple social media platforms.	**AP13:** Collect and test datasets from multiple social media platforms.	N/A
AP14: Develop classifiers that consider multiple targets (i.e., cross-target detectors).	**AP14:** Develop classifiers that consider multiple targets (i.e., cross-target detectors).	N/A
AP15: Document choices related to the online hate classifier development and system functionalities based on the classifier to make these choices transparent and scrutinizable for third parties	**AP15:** Document choices related to the online hate classifier development and system functionalities based on the classifier to make these choices transparent and scrutinizable for third parties	**AP15:** Document choices related to the online hate classifier development and system functionalities based on the classifier to make these choices transparent and scrutinizable for third parties
AP16: Conduct an in-depth error analysis to carefully analyze any systematic errors given by the online hate classifier	**AP16:** Conduct an in-depth error analysis to carefully analyze any systematic errors given by the online hate classifier	N/A
AP17: Share resources, including online hate datasets, algorithms, and models that enable others to replicate results and build upon them.	N/A	N/A
N/A	N/A	**AP18:** Measure and display overall error rates of online hate classifiers to end-users such as moderators.
N/A	N/A	**AP19:** Provide probabilities and explanations of labels provided by the hate classifier for individual samples.

The coverage of this study is limited with focusing on 14 publications. Nonetheless, said publications contained extensive discussions on different challenges of hate detection, and we did observe most of the points we raise here to be repeated across the articles, which implies that our analysis has a degree of saturation, as review works should have. Future work can increase the coverage and depth of the analysis for a more comprehensive survey.

Finally, the analysis could be extended to other stakeholders in further work, e.g., social media companies, legal industry, cyberpsychologists, and so on. For example, the role of the legal perspective in this area is relevant.

V. CONCLUSION

We identified challenges for developing effective online hate detection systems. In terms of future work, stakeholders should obtain datasets that maximize the chances of valid results; detect false positives and minimize the bias associated with the characteristics of the data; disambiguate the semantic context of the messages with attention to hateful words but that in the context do not represent hate (and the opposite situation of obfuscated hate); and respect privacy, anonymity, and freedom of expression when implementing the algorithms. Working on these challenges can yield effective hate detection systems in production.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *Proceedings of Eleventh International AAAI Conference on Web and Social Media*, Montreal, Canada, May 2017, pp. 512–515.

[2] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Community interaction and conflict on the web," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 2018, pp. 933–943.

[3] J. Salminen *et al.*, "Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media," San Francisco, California, USA, Jun. 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17885

[4] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal Attacks Seen at Scale," in *Proceedings of the 26th International Conference on World Wide Web*, Republic and Canton of Geneva, Switzerland, 2017, pp. 1391–1399. doi: 10.1145/3038912.3052591.

[5] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *European Conference on Information Retrieval*, 2013, pp. 693–696.

[6] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," 2011.

[7] H. Almerekhi, H. Kwak, J. Salminen, and B. J. Jansen, "Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions," in *Proceedings of The Web Conference 2020*, Taipei, Taiwan, Apr. 2020, pp. 3033–3040. doi: 10.1145/3366423.3380074.

[8] P. Räsänen, J. Hawdon, E. Holkeri, T. Keipi, M. Näsi, and A. Oksanen, "Targets of online hate: Examining determinants of victimization among young Finnish Facebook users," *Violence Vict.*, vol. 31, no. 4, p. 708, 2016.

[9] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv. CSUR*, vol. 51, no. 4, pp. 1–30, 2018.

[10] J. Salminen, M. Hopf, S. A. Chowdhury, S. Jung, H. Almerekhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Hum.-Centric Comput. Inf. Sci.*, vol. 10, no. 1, p. 1, 2020, doi: https://doi.org/10.1186/s13673-019-0205-6.

[11] B. Vidgen and L. Derczynski, "Directions in Abusive Language Training Data: Garbage In, Garbage Out," *ArXiv Prepr. ArXiv200401670*, 2020.

[12] T. Davidson, D. Bhattacharya, and I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets," in *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, Aug. 2019, pp. 25–35. doi: 10.18653/v1/W19-3504.

[13] S. O. Sood, E. F. Churchill, and J. Antin, "Automatic identification of personal insults on social news sites," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 2, pp. 270–285, Feb. 2012.

[14] S. Modha, T. Mandl, P. Majumder, and D. Patel, "Tracking Hate in Social Media: Evaluation, Challenges and Approaches," *SN Comput. Sci.*, vol. 1, pp. 1–16, 2020.

[15] Y. Senarath and H. Purohit, "Evaluating Semantic Feature Representations to Efficiently Detect Hate Intent on Social Media," in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, 2020, pp. 199–202.

[16] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation (extended version)," *Inf. Syst.*, p. 101584, Jun. 2020, doi: 10.1016/j.is.2020.101584.

[17] P. Mishra, H. Yannakoudakis, and E. Shutova, "Tackling online abuse: A survey of automated abuse detection methods," *ArXiv Prepr. ArXiv190806024*, 2019.

[18] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLOS ONE*, vol. 14, no. 8, p. e0221152, Aug. 2019, doi: 10.1371/journal.pone.0221152.

[19] A. Waqas, J. Salminen, S. Jung, H. Almerekhi, and B. J. Jansen, "Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate," *PLOS ONE*, vol. 14, no. 9, p. e0222194, Sep. 2019, doi: 10.1371/journal.pone.0222194.

[20] S. U. Santuraki, "Trends in the Regulation of Hate Speech and Fake News: A Threat to Free Speech?," *Hasanuddin Law Rev.*, vol. 5, no. 2, pp. 140–158, 2019.

[21] K. J. Madukwe and X. Gao, "The Thin Line Between Hate and Profanity," in *Australasian Joint Conference on Artificial Intelligence*, 2019, pp. 344–356.

[22] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, 2019.

[23] D. Robinson, Z. Zhang, and J. Tepper, "Hate speech detection on Twitter: feature engineering vs feature selection," in *European Semantic Web Conference*, 2018, pp. 46–49.

[24] B. van Aken, J. Risch, R. Krestel, and A. Löser, "Challenges for toxic comment classification: An in-depth error analysis," *ArXiv Prepr. ArXiv180907572*, 2018.

[25] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, "The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2017, pp. 3175–3187. doi: 10.1145/3025453.3026018.

[26] M. Sahlgren, T. Isbister, and F. Olsson, "Learning Representations for Detecting Abusive Language," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, Oct. 2018, pp. 115–123. Accessed: Apr. 23, 2019. [Online]. Available: https://www.aclweb.org/anthology/W18-5115

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv Prepr. ArXiv181004805*, 2018.

[28] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving Google's Perspective API Built for Detecting Toxic Comments," *ArXiv170208138 Cs*, Feb. 2017, Accessed: Jan. 07, 2018. [Online]. Available: http://arxiv.org/abs/1702.08138

[29] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725–726, 2018, doi: 10.1126/science.359.6377.725.

[30] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Hate is Not Binary: Studying Abusive Behavior of #GamerGate on Twitter," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, New York, NY, USA, 2017, pp. 65–74. doi: 10.1145/3078714.3078721.

[31] M. Castelle, "The Linguistic Ideologies of Deep Abusive Language Classification," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, Oct. 2018, pp. 160–170. Accessed: Apr. 22, 2019. [Online]. Available: https://www.aclweb.org/anthology/W18-5120

[32] Y. Kou and B. Nardi, "Regulating anti-social behavior on the Internet: The example of League of Legends," in *Proceedings of iConference*, 2013, pp. 616–622. doi: 10.9776/13289.