The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection

Alexandra Olteanu
IBM Research

Kartik Talamadupula
IBM Research

Kush R. Varshney
IBM Research

ABSTRACT

Wagstaff (2012) draws attention to the pervasiveness of abstract evaluation metrics that explicitly ignore or remove problem specifics. While such metrics allow practitioners to compare numbers across application domains, they offer limited insight into the impact of algorithmic decisions on humans and their perception of the algorithm's correctness. Even for problems that are mathematically the *same*, both the *real-cost* of (mathematically) identical errors, as well as their *perceived-cost* by users, may significantly vary according to the specifics of each problem domain, as well as of the user perceiving the result. While the real-cost of errors has been considered previously, little attention has been paid to the perceived-cost issue. We advocate for the inclusion of *human-centered metrics* that elicit error costs from humans from two perspectives: the nature of the error, and the user context. Focusing on hate speech detection on social media, we demonstrate that even when fixing the performance as measured by an abstract metric such as *precision*, user perception of correctness varies greatly depending on the nature of errors and user characteristics.

KEYWORDS

Evaluation metrics; hate speech; human-centered metrics

1 ON THE USE OF ABSTRACT METRICS

In computing research the goal is often to develop methods that improve on some well-established and well-understood metrics. Such metrics are typically abstract and domain-agnostic to allow comparisons across different settings and domains [6], such as the precision of a classification method. As a result, their real-world impact–"*dollars saved, lives preserved, time conserved*" [6]—is often unknown and may diverge depending on the specifics of each domain [3]. The same point can be made about how well such metrics estimate users' *perception* of performance, which may further vary according to their idiosyncrasies and context. For instance, White [7] reports that when searching for medical information, users often settle on wrong answers that confirm their personal beliefs. This is concerning, suggesting that evaluations that optimize for domain-agnostic metrics may also need to correct and account for what users may *perceive* as the correct or wrong answer.

WebSci '17, June 25-28, 2017, Troy, NY, USA

ACM ISBN 978-1-4503-4896-6/17/06.
https://doi.org/10.1145/3091478.3098871

We build on the argument of Wagstaff that *abstract metrics* are insufficiently grounded in the application domain to offer meaningful insights about the effects of the performance numbers that they output for a given domain. However, focusing on the cost of different types of errors, we distinguish between *real-costs* that are directly measurable—such as response time or monetary gains—and *perceived-costs* that attempt to capture what the cost of an error *appears to be* for the users of a system. While these costs are often contingent on each other or may even overlap, there may also be important differences in the signals they capture.

Error Costs Vary Across & Within Domains. For discussion purposes, let us consider the task of classifying content from social platforms. Even in this limited setting, achieving a certain accuracy, e.g., 65%, might be adequate for some applications (e.g., identify cat pictures for image search), but not for others (e.g., identify online hate speech for content moderation) [4, 6]. Thus, given a task to be algorithmically performed and a fixed performance threshold, there may be wide variations in how the same performance on a given metric translates in terms of both real-costs, and *how adequate users perceive the performance of the system to be*. Even with an overall high performance for a classification task and application domain, it is hard to know what that implies, as it might also vary widely across the feature space [2]. We may see a high performance for one category of users or data, and a low one for a different category that may represent a minority; or we may even see high cost variations for errors associated with different classes of users.

On the Inclusion of "Human-in-the-loop" Metrics. On social platforms, algorithmic decisions are typically made *for* people (e.g., search results) and/or *about* people (e.g., friend recommendations) [5]; and user participation is important for both of these tasks. This should have made the use of a standard set of *human-centered metrics* across semantically different domains commonplace; alas, this is rare. Such *human-centered metrics* should account, among others, for how the *perceived-cost* of erroneous decisions varies depending on (1) the specific problem (e.g., hate speech detection); (2) the type of the mistakes; and (3) user characteristics and context.

In this abstract, we advocate for research into characterizing how the perceived and the real costs of different kinds of errors—and the utility of popular metrics—depend on the problem domain specifics, and on the peculiarities of users that evaluate and are affected by those errors. Through a brief case study on hate speech detection on social media we provide initial evidence into how the perception of performance is influenced by users' context and the type of errors an automated system makes.

2 CASE STUDY: HATE SPEECH DETECTION

To quantify the relation between abstract metrics and user perception of correctness, in our experiment we set the performance of a classification system and vary the nature of the mistakes it makes

to measure their impact on the users' overall perception of quality. We also measure how user traits appear to affect their evaluations. We present preliminary results from our experiments, where the application domain is *hate speech detection on social media.*

Dataset & Experimental Setup. We use the initial dataset released by [1], which contains over 14*K* tweets annotated as *contains hate speech* (2399 tweets), *uses offensive language but not hate speech* (4836 tweets), and *is not offensive* (7274 tweets); allowing us to distinguish between different kinds of *false positive* errors that an automated hate speech detection system for social media may make: classifying as hate speech messages containing offensive language but not hate speech, versus classifying as hate speech messages that neither contain hate speech nor are offensive. For our evaluation, to avoid borderline cases, we filter out those tweets whose labels were established with a low confidence (< 70%).

Using this dataset, we design a crowdsourcing task that asks annotators on the GetHybrid.io platform to evaluate an (hypothetical) automated system that outputs social media messages it classifies as *hate speech*. We generate random samples of 8 messages each, containing from 0 to 7 misclassified instances (corresponding to precision scores varying from 100% to 12.5%); instances sampled either from messages containing offensive language but not hate speech (*low cost errors*), or from messages that are not offensive (*high cost errors*; as the distinction between these messages and hate speech message should be clearer). For each precision score and type of error, we generate 6 different samples (to ensure results are not an artifact of a given sample), and for each of them we collect annotations from 5 different annotators. This resulted in 30 annotations for each precision score. Finally, to understand how user traits may affect their evaluation, we also ask the annotators if they have been the target of hate speech, and if they believe hate speech should be moderated on social media—see Table 1 for the exact questions and the distribution of answers.

Does the nature of errors influence users' perception of quality? Figure 1 (left plot) shows how users' perception of quality varies as the system makes more errors (the precision decreases) when the "misclassifications" are sampled from messages that are offensive but not hate speech (*low cost errors*), versus when the sampling is done from messages that are not offensive (*high cost errors*). We see that the perception gap between the two cases increases with the error rate—showing that classifying too many non-offensive messages as hate speech leads to a more rapid drop in quality perception, while for misclassification of other types of offensive language it flattens. This pattern holds even if we condition on the users experience with *hate speech*, or their stance on online moderation.

Do user attributes play a role in how they evaluate quality? Figure 1 (right plot) shows a consistent gap in evaluations between annotators that self-identify as targets of hate speech vs. those that report never having been a target of online harassment. Noting first that annotators who never experienced online harassment make more accurate evaluations w.r.t. the original annotations (figure omitted), we observe that the evaluation gap among the two groups tends to widen mainly for *low cost errors*. It appears that those that were the target of hate speech apply a broader definition of what constitutes hate speech. We noticed similar patterns between those

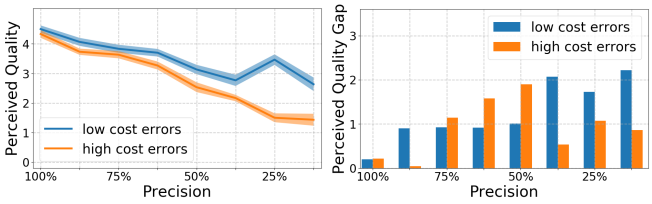

Figure 1: Quality perception variation as a function of precision. Low-cost errors refer to offensive tweets classified as *hate speech*, while high-cost errors to non-offensive tweets classified as *hate speech*. Left plot: quality perception for errors of different kinds. Right plot: differences in quality perception between annotators targeted by hate speech vs. those that never experienced online harassment.

Total	**480**
Were you ever the direct target of hate speech?	
Yes, unfortunately	213
No, but I've experienced other forms of online harassment	176
No, never	91
How important it is to moderate hate speech content on social media?	
I think this is increasingly necessary.	104
Some form of moderation is needed, but I also worry about free speech rights.	373
It is not necessary. If you don't like what folks say, do not engage with them.	3

Table 1: Answers distribution according to annotators experience with "hate speech" and their stance on moderation

that think there is an increasing need to moderate hate speech online versus those that take a more moderate stance on this, with the latter again making more accurate evaluations. Untangling the relation between the two self-declared traits and its impact on user evaluations is left to future work.

3 CONCLUSIONS

Our results are preliminary and larger scale experiments across multiple application domains are needed. However, we believe they provide initial evidence that (1) the nature of errors made by automated systems affects users' overall perception of quality, and (2) user traits also play a role (e.g., being a victim of hate speech or not, or their stance on online moderation). This supports our main point of wanting to elicit quality metrics based on the semantics of the problem rather than allowing them to be domain-agnostic.

REFERENCES

[1] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proc. of ICWSM.*

[2] Moritz Hardt. 2014. How big data is unfair: Understanding sources of unfairness in data driven decision making. *Medium. Available at: https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de* (2014).

[3] Pat Langley. 2011. The changing science of machine learning. *Machine Learning* 82, 3 (2011).

[4] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *SSRN Preprint* (2016).

[5] Kush R Varshney. 2015. Data science of the people, for the people, by the people: A viewpoint on an emerging dichotomy. In *Bloomberg Data for Good Exch. Conf.*

[6] Kiri Wagstaff. 2012. Machine learning that matters. In *Proc. of ICML.*

[7] Ryen White. 2013. Beliefs and biases in web search. In *Proc. of SIGIR.*