Hate Speech: a Quality of Service Challenge

Andre Oboler
Online Hate Prevention Institute
Melbourne, Australia
Oboler@ieee.org

Karen Connelly
Cyber Racism and Community Resilience Project
University of Technology Sydney
Sydney, Australia
Karen.J.Connelly@student.uts.edu.au

Abstract—**The use of social media has become pervasive across many aspects of our lives. We now depend on social media platforms more than ever before. Our dependence on social media has created a greater demand for a higher quality of service, but also for this quality to apply in new areas. As the social media experience and "real world" experience merge, there is an increased expectation that the norms of society will also apply in social media settings. There is an increasing demand for social media platforms to empower users with tools to report hate speech and other forms of dangerous content. There is also an increasing demand for greater quality of service in the way these reports are managed. The approach of social media companies to this problem, which is to largely avoid the issue by not publishing the data needed to assess the relevant quality of service, is being overcome by third party solutions. This paper discusses one such solution which is currently under development, as well as some of the challenges to improving quality of service in this area.**

e-services; quality of service; hate speech; social media

I. Introduction

E-Services are services delivered to businesses and customers through the internet [1]. They can involve individual standalone services, or compositions made of multiple e-services potentially coming from multiple suppliers [1]. Social media platforms, such as Facebook and Twitter, provide e-service both directly and by facilitating the inclusion of third party e-services.

Quality of Service (QoS) is a non-functional requirement of a system [2]. In an e-services system, service availability and response time are important factors in determining whether the system is fit for the purpose, however, other factors such as compliance with privacy requirements can also form part of QoS [2]. Expectations across a range of requirements can be outlined in a QoS policy for an e-Service [2].

Social media companies have ensured their platforms are built for growth. The platforms are built to handle large numbers of users, provide fast response times, and minimize or eliminate service unavailability. Other aspects of QoS, such as the degree to which users have complied with the platforms policies, have been largely forgotten. Facebook, for instance, estimated that 8.7% of its user profiles were fake accounts and only began to seriously tackle the problem in 2012 [3]. There is a growing push for improved QoS in social media, as measured across multiple dimensions, as a result both of the increased significance of social media and its transformation into a mainstream part of daily life.

This paper begins with a discussion of the need for greater QoS in social media as user expectations grow, and governments seek to change the regulatory environment in which social media companies operate. We focus specifically on the pressure for greater QoS in the handling of hate speech.

Next we examine the difficulties social media companies face in handling hate speech. A number of challenges are outlined, including the need to recognize local variations of racial slurs and other forms of hate speech. The platforms solution so far has been not to release the data which would be needed to evaluate the scope of the problem, and the quality of the platforms response.

The paper goes on to discuss a new solution, currently being built, which seeks to independently collect and make available data on the scope of the hate speech problem and the QoS of the platforms' responses. We conclude by noting how the new solution could prove to be a game changer in making QoS measurable and social media companies more publicly accountable when their actions greatly differ from societal expectations.

II. The Need for Quality of Service in Social MEdia

Social media services, which were once part of the Wild West of the Internet, are now part of everyday suburbia. The transformation from the latest fad into a part of mainstream life leads to new expectations on social media platform providers from both citizens and governments. New expectations alter the QoS requirements for social media platforms, and so far platforms have remained largely unresponsive until pushed by external pressures.

A. *The Significance of Social Media*

Social media is pervasive. The latest data from PEW Research indicates that 84% of American adults are online, and of these 74% use at least one social media platform [4, 5]. Usage is not uniform: 71% of online American adults use Facebook, 22% use LinkedIn, 21% use Pintrest, 18% use Twitter and 17% use Instagram [6]. While the volume of use shows social media is a part of most people's lives, the efforts of governments to keep social media online, or conversely to block access to it, shows the perceived impact of social media on the lives of people and nations.

In 2009, as protesters took to the streets following the Iranian elections, a US State Department official e-mailed one of the Twitter co-founders to request that scheduled maintenance, which would have temporarily taken the micro-blogging site offline, be deferred [7]. The US State Department's view that Twitter was "playing an important role at a crucial time" was seen as a milestone by the New York Times who described it as "recognition by the United States government that an Internet blogging service that did not exist four years ago has the potential to change history in an ancient Islamic country" [7]. That same year, both Facebook and Twitter were blocked by the Chinese Government following an outbreak of violent riots [8]. In the 2011 Egyptian elections the United States Government sought to keep social media online, while the Egyptian Government sought to prevent its use. In that instance the Egypt Government ultimately prevailed by cutting Internet access for the entire country [9].

The impact of social media on our lives continues to grow, as does our reliance on social media for personal communications and access to news, education and employment. Companies rely heavily on social media for branding and marketing. Social media has become embedded in our daily lives. At the same time, social media poses risks to community cohesion, public safety, and individual rights and freedoms. This creates a need for social media companies to operate in a predictable manner which meets the expectations of society as expressed directly by citizens, and by the governments which represent them. There is in short, a demand for a certain quality of service as measured across multiple dimensions.

B. The Demand for Quality of Service

The danger of hate speech in social media, and its ability to make racism social acceptable, was first raised in 2008 [10]. Calls have been made since at least 2010 for companies that profit from the communication they facilitate in social media to assume public obligations with respect to the content they carry. Two obligations that have been suggested are an obligation to "take reasonable steps to discourage online hate" and an obligation see such content "removed within reasonable time" [11]. Both concepts imply a measure of a service which can be compared to a standard. This approach can be viewed as one based on quality of service, with the potential for minimum standards imposed through legislation.

There is also an intrinsic need for quality of service with respect to the removal of hate speech. Jäkälä and Berki have noted that online communities have "expression boundaries as well as norms and rules for behavior on-line and sometimes also off-line". They highlight that "cyberspace does not exist without electronic inhabitants; otherwise it is a deserted cyber place", and that at the heart of a successful online community lies a sense of "belonging, mutual respect, and commitment" [12]. A failure in the enforcement of expression boundaries, rules and norms, such as Facebook's "Community Standards", is likely to impact on users commitment. Where the lack of enforcement occurs with respect to rules designed to protect users' sense of belonging and mutual respect, this degrades the quality of the online community which is at the core of the service.

Like other communications providers, social medial platforms may also find themselves subject to imposed standards of quality due to legislation. Governments have expressed concern over the quality of report handling by social media companies [13]. A consultation paper from the Australian Government in 2014 stated that "Australians currently have no recourse in instances where they may disagree with how their content complaints are handled by social media sites" [13]. In response the Government expressed an intention to legislate for a new scheme "to enable the rapid removal from a large social media site of material targeted at and likely to cause harm to a specific child" [13].

Even where standards can't be imposed by law, for example in the United States, steps are being taken to increase the pressure on Social Media companies. The National Science Foundation, for example, has allocated a million dollars to a project that aims to create a monitoring service for Twitter. The service will "mitigate the diffusion of false and misleading ideas, detect hate speech and subversive propaganda, and assist in the preservation of open debate" [14].

III. The Challenege of Hate Speech

Managing hate speech poses a particular challenge to social media companies. The challenge is in part cultural, as most social media companies are based in the United States and are embedded in a First Amendment culture which sees hate speech as a form of speech protected from government regulation. There is therefore a reluctance to remove such content even when the platforms own policies prohibit hate speech. There is also a lack of local case law, or public understanding, on when content has crosses into hate speech.

Hate speech has been defined as 'speech or expression which is capable of instilling or inciting hatred of, or prejudice towards, a person or group of people on a specified ground including race, nationality, ethnicity, country of origin, ethno-religious identity, religion, sexuality, gender identity or gender' [15]. While there are many forms of hate speech, racism is the most general and best regulated form of hate speech. Racist speech relates to a socially constructed idea about differences between social groups based on phenotype, ancestry, culture or religion [16]. This section will largely focus on racism, although the challenges apply, often to an even greater extent, to other forms of hate speech as well.

A. The need for regulation

The push to regulate hate speech is in part a result of globalization with the values of other countries, and of Europe in particular, playing a larger role in online culture. There is also a growing body of scientific evidence which highlights the real impacts on health that results from racism. Where companies are left to make up their own mind on whether to regulate for hate speech or not, this evidence provides a persuasive argument.

A systematic review of empirical research on self-reported racism and health, for example, found a strong correlation between racism and negative mental health outcomes such as depression, anxiety and emotional stress and also some association between racism and negative physical health outcomes such as high blood pressure and low infant birth weight [17]. Similarly, a review of the health effects of racism on children and youth found a strong relationship between racism and negative mental health outcomes including anxiety, depression and low self-esteem [18].

A recent study on the relationship between online victimization and psychosocial problems in adolescents found "convincing evidence" that online victimization causes feelings of loneliness and social anxiety [19]. Another study exploring online racial discrimination and psychological adjustment found that "consistent with offline studies, online racial discrimination was negatively associated with psychological functioning" [20]. The negative health impacts of racism, including online, cause significant social stress. Providing such an environment will not be in the interests of social media platforms.

B. *Recognizing it when it occurs*

Racism is often expressed though negative and inaccurate stereotypes, emotions such as fear and hatred and actions such as threats, insults and discrimination embedded in social systems and structures [18]. Even if it wants to intervene, for a social media platform to correctly respond to a complaint about racism, the person assessing the complaint would need to understand the message being communicated, and would need to recognize whether that message was racist. This may be a difficult challenge requiring specific background knowledge.

The knowledge required to recognize racism from a particular location may also require knowledge of the local context. Research in Europe has found that in countries where the majority of the population are Christian, people from other religions, such as Muslims are likely to be seen as outsiders [21]. In Australia there have been difficulties in getting social media companies to recognize "Aboriginal Memes" targeting Indigenous Australians as hate speech [22]. There is also the problem of local derogatory slang, or the use of images which require local knowledge to be understood.

Legal differences also create difficulty. Social media platforms often seek to block content within a jurisdiction where the content is unlawful rather than removing it from the platform as a whole. If content is not seen to breach the platforms own terms of service, it may still need to be assessed based on the law in the jurisdiction of the person who reported it. This could lead to the same content having to be assessed multiple times according to multiple different laws. Just within Australia, different forms of hate speech are unlawful in different states [23].

IV. FIGHT AGAINST HATE: A STEP TOWARDS A SOLUTION

A. *Origins of Fight Against Hate*

The Fight Against Hate software was initially designed to meet the need for data as outline by the Online Antisemitism Working Group of the Global Forum to Combat Antisemitism (GFCA) in 2009 [24]. The working group, made up of experts from around the world, highlighted the need for metrics on:[24]

- the number of problem items per platform
- the number of items resolved per platform
- the time between something being reported and action being taken

This need for data was reiterated by the Inter Parliamentary Coalition to Combat Antisemitism in their London Declaration which called for an "international task force of Internet specialists comprised of parliamentarians and experts… to create common metrics to measure antisemitism and other manifestations of hate online" [25]. The London Declaration highlighted the need for this data to be gathered in terms of all forms of online hate.

In 2011 a software project specification for Fight Against Hate was presented to the Online Antisemitism Working Group of the GFCA, and at the International Conference in 2013 an action item was recorded to proceed with the creation of a database to record antisemitic websites and social media content [26]. This was endorsed as a major focus of the GFCA as a whole by the Steering Committee in 2014. The intention is to deliver major announcements in relation to the software at the International Conference of the GFCA in 2015.

The origins of the software have largely determined its requirements and ensured significant stakeholder involvement, particularly from governments and non-governmental organizations.

B. *The Approach: Crowdsourcing and validation*

Recognizing that hate speech can be difficult to find, particularly when the content is embedded in images or video, or when it is heavily dependent on context, the approach adopted was to start with assessments made by users. The software therefore provides a mechanism where users can report online content. Users are asked to only report content they have first reported through a platforms own reporting mechanism.

The system captures not only the details of the content a user reports, but also when the report was lodged. The system also maintains a list of content each user has reported and allows users to indicate if the content has been taken offline. The time a user reports content as being removed is also recorded. The system maintains a record of the first report of any new item of content, and the first report of it being removed, and from this data a measure of a minimum time that the content remained online can be calculated.

Users report content into the system by logging in and then providing a he URL of the problematic content. Following this the user is asked to classify what kind of hate speech they are reporting, and how confident they are in their classification. For some hate types a second stage of classification asks for a sub-classification and level of confidence.

Users operate as part of team and can see the reports filed by other members of their team. This allows them to report those items to the platform concerned, increasing the chance of a positive outcome. A point system provides some gamification allowing users to see the contribution they are making to the task of monitoring online hate and to the impact of their own team.

The major weakness of this system is users, either maliciously or through ignorance, misreporting content. We expect there will be significant false positives in the system, as well as efforts by organized groups to misuse the software to report political content they disagree with or rivals of any type (e.g. sporting teams, businesses, etc). To address this concern, users of the system will also be encouraged to review items of content which do not come from their team. A button is provided to request an item, and items will then be allocated. This allows the opinion of multiple randomly selected users to be compared against those who self selected in reporting the content.

Further quality control is provided using experts who assess and verify a limited volume of reports. These reports can then be used to assess how effective users are in their reporting. This validation is similar to that used in assessing artificial intelligence agents, only in this case the agents are real users. In this way a number of users with known proficiency can be gathered and then used to validate unknown items and check the quality of other users.

C. Public Transparency

Through the collection of a significant volume of data it will become possible to calculate average response times across different platforms and for different types of hate speech. It will also become possible to assess the accuracy of social media systems in identifying online hate. The volume of data on hate speech against each group in society, as seen from each country, is also of significant interest. This data will help improve the transparency of social media services.

Once there is more transparency on the scale of the problem in social media, and on the response being provided (or not provided) by large scale social media companies, it is reasonable to expect that improvements in QoS will follow. This is a primary goal of the system, to make the platforms themselves publicly accountable.

D. An Expert Tool

Additional features in the software provide a service for researchers, civil society organizations and government agencies. This service allows experts to see all the hate items reported by people from within the expert’s geographic area of interested and matching the hate classification the expert is interested in. In a later phase of development experts will also be able to specify a minimum degree of confidence the system must have in the item meeting the expert’s criteria before the expert sees it.

This access to individual items that have been reported will allow experts to carry out their own research into the nature of online hate, and to track spikes and trends in the data. It will also allow relevant agencies to take action to have items removed by platform providers, or blocked in their local jurisdiction if such content violates national laws.

Further tools in a late stage of development will allow experts to review items that have been in the system beyond a specified minimum number of days, and which have been recently verified to still be online. This data will allow experts to find content which users consider hate speech, but which platforms are failing to remove. Research into such incidents can help to address limitations negatively affecting QoS.

The data discovered by these expert tools may also indicate failures of QoS policies, some of which may be backed by fines or other government imposed penalties. The tool can therefore aid in the monitoring of compliance with agreed or imposed QoS standards.

E. The Technology

Fight Against Hate has been created as e-Service built on Amazon Web Services (AWS) technology and programmed primarily in PHP. The database uses DynamoDB, a NOSQL database which allows significant scale. The database design allows additional hate types to be added without restructuring the database.

Fight Against Hate reuses and stores IDs from YouTube and Facebook simplifying the task of tracking and loading reported content in the original platform. The data may also be of use to law enforcement agencies wanting to request IP addresses related to specific content.

While currently running on our own service, the intention is for the software driving the application to also be migrated to the cloud, and specifically onto an EC2 instance, an AWS virtual server. This will allow the processing needs to be scaled up as the volume of use expands.

F. Meeting the challenge

We previously discussed the difficulty of staff at a social media company being versed in all forms of online hate. By relying on the crowd, and allowing users who review content to be selected by the system, Fight Against Hate is able to access both local knowledge and the expertise of particular users. Such an approach is likely to provide more accurate results than the methods platform providers currently use.

By allowing many different experts to use the system, it is also more likely that a local expert will come into contact with data that requires local knowledge. This in turn can lead to reports and publications which improve public understanding of local forms of hate speech.

Both approaches have the potential to create a feedback loop which improve the understanding of staff at social media companies, ultimately leading to better initial assessment, and more hate speech being removed.

V. Conclusions

This paper highlights the need for Quality of Service to be considered more holistically when it comes to social media. There are increasing expectations from both users and governments when it comes to the removal of hate speech from social media platforms. The process of assessing and removing hate speech is a fundamental part of the QoS in such platforms. There is a significant risk that a failure to improve QoS will lead to minimum standards being imposed.

So far the imposition of minimum standards has been avoided largely due to lack of data to support legislation. Platform providers have avoided sharing data on the magnitude of the online hate problem, or on the quality of their efforts in response. As public pressure builds it appears regulation may be imposed even without a base of evidence to support such changes.

The paper outlined some of the challenges platforms face in identifying hate speech. It also presented a new tool, currently in development, which uses the crowd to address some of these difficulties. The tool presented will create summary statistics which can help to assess the QoS of platform providers efforts in removing hate speech, and can also be used to identify cases quickly that fail to comply with QoS policies. The platform will also provide a valuable tool for research into online hate, allowing the problem to be addressed more effectively in the future.

References

[1] F. Casati and M. Shan, “Dynamic and adaptive composition of e-services,” *Information systems*, vol. 6, no. 3, 2001, pp 143—163.

[2] A. Slimane and C. Souveyet, “A Goal Driven Approach to Deal with Quality of Service as Potential Aspects,” in *International Workshop on Advanced Information Systems for Enterprises*, Constantine, Algeria, 2008, pp 18—24.

[3] E. Protalinski, “Facebook: 8.7 percent are fake users”, *CNet News*, Aug. 1 2012. [Online]. Available: http://www.cnet.com/news/facebook-8-7-percent-are-fake-users/

[4] Pew Research Internet Project, “Internet User Demographics.” [Online]. Available: http://www.pewinternet.org/data-trend/internet-use/latest-stats/. [Accessed: Aug. 30 2014].

[5] Pew Research Internet Project, “Social Media User Demographics.” [Online]. Available: http://www.pewinternet.org/data-trend/social-media/social-media-user-demographics/. [Accessed: Aug. 30 2014].

[6] Pew Research Internet Project, “Social Media Update 2013.” [Online]. Available: http://www.pewinternet.org/2013/12/30/social-media-update-2013/. [Accessed: Aug. 30 2014].

[7] M. Landler and B Stelter, “Washington Taps Into a Potent New Force in Diplomacy,” *The New York Times*, Jun. 16 2009. [Online]. Available: http://www.nytimes.com/2009/06/17/world/middleeast/17media.html?_r=0

[8] R. Wauters, “China Blocks Access To Twitter, Facebook After Riots,” Tech Crunch, Jul. 7 2009. [Online]. Available: http://techcrunch.com/2009/07/07/china-blocks-access-to-twitter-facebook-after-riots/

[9] A. Oboler et al., “The danger of big data: Social media as computational social science” *First Monday*, vol. 17, no. 7, Jul 2012. Available: http://firstmonday.org/article/view/3993/3269/

[10] Oboler, A., “Online Antisemitism 2.0. Social Antisemitism on the Social Web”, *Jerusalem Center for Public Affairs Post-Holocaust and Antisemitism Series*, no. 67, Apr. 2008. [Online]. Available: http://jcpa.org/article/online-antisemitism-2-0-social-antisemitism-on-the-social-web/

[11] Oboler, A., “Time to Regulate Internet Hate with a New Approach?” *Internet Law Bulletin*, vol. 13, no. 6, pp 102—106, Oct 2010.

[12] M. Jäkälä and E. Berki, “Communities, Communication, and Online Identities,” in *Digital Identity and Social Media*, S. Warburton and S. Hatzipanagos, Eds. Pennsylvania: IGI Global, 2013, pp. 1—13.

[13] Australian. Department of Communications, *Enhancing Online Safety for Children: Public consultation on key election commitments*. Jan. 2014. [Online]. Available: http://www.communications.gov.au/__data/assets/pdf_file/0016/204064/Discussion_Paper_-_Enhancing_Online_Safety_for_Children.pdf

[14] E. Harrington, “Feds creating database to track hate speech on Twitter”, *Fox News*, Aug. 26 2014. [Online]. Available: http://www.foxnews.com/politics/2014/08/26/feds-creating-database-to-track-hate-speech-on-twitter/

[15] K. Gelber and A. Stone, “Introduction,” in *Hate speech and freedom of speech in Australia*, K. Gelber and A. Stone, Eds. Sydney: The Federation Press, 2008, pp xiii.

[16] Y. Paradies, “Defining, conceptualizing and characterizing racism in health research”, *Critical Public Health*, vol. *16,* no. 2, pp. 143—157, Jun. 2006.

[17] Y. Paradies, “A systematic review of empirical research on self-reported racism and health,” *Int. J. of Epidemiology*, vol. *35,* no. 4, pp. 888–901, Apr. 2006.

[18] N. Priest et. al., “A systematic review of studies examining the relationship between reported racism and health and wellbeing for children and young people,” *Social Science & Medicine*, vol. *95*, pp. 115—27, Oct 2013.

[19] R. Van den Eijnden et. al., “The bidirectional relationships between online victimization and psychosocial problems in adolescents: a comparison with real-life victimization” *J. of Youth and Adolescence*, vol. 43, no. 5, pp. 790—802, Aug 2013.

[20] B. Tynes et al., “Online racial discrimination and psychological adjustment among adolescents,” *J. of Adolescent Health,* vol. 43, no. 6, pp. 565—569, Aug. 2008.

[21] A. Zick, B. Küpper, and A. Hövermann, *Intolerance , Prejudice and Discrimination - A European Report.* Berlin: Friedrich-Ebert-Stiftung, 2011.

[22] "Facebook shuts vile Aboriginal memes page, despite earlier claiming it didn't constitute 'hate speech'", *News.com.au*, Jan. 27 2014. [Online]. Available: http://www.news.com.au/technology/facebook-shuts-vile-aboriginal-memes-page-despite-earlier-claiming-it-didnt-constitute-hate-speech/story-e6frfrnr-1226811373505

[23] A. Oboler, “Legal Doctrines Applied to Online Hate Speech,” *Computers & Law*, no 87, pp 9—15, July 2014.

[24] A. Oboler and D. Matas (Eds), “Report from the Working Group on Online Antisemitism”, in *Global Forum to Combat Antisemitism*, Jerusalem, Israel, 2009.

[25] Inter-Parliamentary Coalition for Combating Antisemitism, *London Declaration on Combating Antisemitism*. Feb. 17 2009. [Online]. Available: http://www.antisem.org/london-declaration/

[26] “The Action Plan for Combating Antisemitism 2013 and Beyond”, *The 4th International Conference of the Global Forum for Combating Antisemitism*, Jerusalem, Israel, May 2013. [Online]. Available: http://mfa.gov.il/MFA/AboutTheMinistry/Conferences-Seminars/Documents/AntisemitismBooklet2013.pdf