An analysis of homophobia on vandalism at Wikipedia

Carlos Alberto Damas
Institute of Computing – UFF
Niterói, Brazil
carlosatd@id.uff.br

Karina Mochetti
Institute of Computing – UFF
Niterói, Brazil
kmochetti@ic.uff.br

***Abstract*—Wikipedia is currently one of largest source of digital information. With billion page views per month worldwide, the free digital encyclopedia is based on collaborative editing, in which any user, even anonymously, can edit and include content. Although such content is always evaluated by several users daily and also by more experienced ones, vandalism is still common, especially on pages with sensitive information. In this paper we evaluated the vandalism found on pages of LGBT computer scientists, comparing them with scientists who are not part of this minority and analyzing the types of vandalism found. Thus, this paper analyzes how being part of a minority can influence relationships in the digital society.**

***Index Terms*—Wikipedia, Vandalism, Equity, LGBT**

I. INTRODUCTION

Wikipedia is the biggest encyclopedia created and works based only on collaboration, encouraging its readers to submit and review content on its own articles. This has lead to a respectable number of 13 millions registered users in 240 different languages [1]. As it may be expected from its powerful magnitude, although having rules and guides, it is still common that some people misuse it deleting entire pages, or adding aggressive and offensive content. This actions are considered vandalism according to Wikipedia's guidelines (https://en.wikipedia.org/wiki/Wikipedia:Disruptive_editing).

Identifying vandalism in Wikipedia proves itself to be a hard task due to lesser cases in which newcomers submit non-intentionally and non-offensive content [2]. In order to provide a better approach, Wikipedia itself has generated a great amount of content and policies about vandalism, summarizing the most commons mistakes made by new editors. In that way, it tries to clear the thin line between mistakes and malicious attack. This leads to a list of 20 most common vandalisms found on Wikipedia [3]. From this, we highlight four:

- **Blanking:** Deleting parts of a page or even all of its content, without a reasonable explanation.
- **Adding Ill Intent Content:** Hoaxing, adding spam, non-relevant very large content, profanity, graffiti, nonsense, defamatory or slanderous to an article.
- **Hidden Vandalism:** Adding ill intent content, only readable during edition or inspection of the page, as using embedded text or link editing.
- **Page Forms Editing:** Changing the page's name, template, images, format, summary with ill intentions.

While most researches focus on detecting and removing vandalism, few aim on its content and intentions. The behaviour of Wikipedia trolls can be compared to hackers which find pleasure only from causing damage [4]. Although being a digital environment, Wikipedia is a community that can reflect our culture and our society [5]. Therefore the quantity, content and motivation behind those vandalisms can culturally impact the inclusion of minorities, such as LGBT+, on some fields, such as Computer Science.

Computer Science field, as well as the Wikipedia community, are known to be a male-dominated environment in which minorities are hardly represented. In this work we intent to relate the LGBT+ representation of Wikipedia and Computer Science by analyzing the vandalism on computer scientists pages, showing its hostile environment toward minorities.

II. METHODOLOGY

Our focus in this work is to categorize and analyze vandalisms in pages that are part of LGBTQ+ community specifically from the Computer Science field. In order to do so we will use tools that are freely available to all registered users through the Wikimedia Foundation Projects.

With that purpose we use the following tools: Quarry is a public querying interface that grants access to Wiki Replicas, a set of SQL database, allowing users to fetch different columns and rows as data [6]. ORES (Objective Revision Evaluation Service) is a web service designed to provide machine learning as a service to Wikimedia Projects, evaluating pages qualities, for example [7]. To reach higher levels, such as *Good Article* and *Featured Article*, a page must be nominated and evaluated by one or more impartial reviewers. Each page is then associated with a level: from 1 for *stub* to 7 for *Featured Article*.

Other tools used include PAWNS, Wikimedia's personalized Jupiter notebooks and Python libraries such as JSON [8] and PANDA [9] to organize and analyze the data.

With the aid of these tools, we were able to generate the quality graphic of Wikipedia pages and detect vandalisms. An example code can be found at the author's Github page (https://github.com/Carlosatd/VandalismCode).

III. RESULTS

To better select our candidates we have looked for openly LGBT+ within the Computer Science field. We used different search criteria as well specific pages such as 'LGBT

Scientists', 'LGBT People by occupation and nationality' and 'Transgender and transsexual computer programmers'. Finally, we sought Non LGBTQ+ Scientists pages in order to compare not only the number of vandalisms but also its contents and relevance. It is important to notice that women, African Americans and other minorities were excluded from this category. We would like to compare LGBTQ+ minority with a group not included in any other minority, so our results can be clearly attributed to the participation of the scientist in this specific group and no other.

And so we ended with the names that follow:

- Openly LGBT+ Scientists: Alan Turing, Peter Landin, Tim Gill, Luca Trevisan, Edith Windsor, Sophie Wilson, Audrey Tang, Lynn Conway.

- Non LGBT+ Scientists: Bill Gates, David Patterson, Andrew Tannenbaum, Guido van Rossum, Tim Berners-Lee.

Figure 1 shows the quality graphic found on Alan Turing and Bill Gate's pages. Alan Turing was an English computer scientist known as the father of theoretical computer science. He was convicted by United Kingdom law for being homosexual in 1952. His page was the only GA level found and, therefore, it has the largest number of vandalisms. A few of them were mostly harmless ones but the bigger part used words and slangs to directly offend using his sexual orientation as a motif.

Bill Gates is the main founder of Microsoft Corporation. His page has a lot of information, reaching an GA level and by this being almost completed. It is semi-protected since 2011 due excessive vandalism that went from simple blanking, nonsense addition to harsher offensive slangs, a lot with sexual connotation mostly directly using gay as an offense.

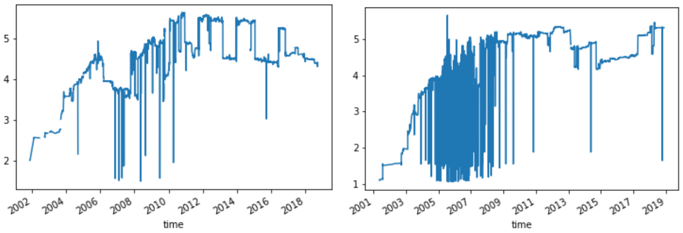

Fig. 1. Quality graphic for Alan Turing's page, on the left and Bill Gate's page on the right.

The first fact to notice is the lack of LGBT+ Computer Scientists on Wikipedia. Although we are taking into account only the openly LGBT+ scientists, this representation gap is still an issue, rather being for the lack of space to LGBT+ on the field or for the lack of encourage to computer scientists to come clean about their sexual orientation.

By analyzing the data we could notice a difference between pages level score of those LGBT+ notable computer scientists with mostly being C, while non-LGBT+ have at least a level B article in most cases. Since vandalisms occur more often on popular pages, it will be common to find more vandalisms on articles with higher level as we did.

It is important to notice, though, that when vandalism is found on LGBT+ pages, it usually contains insults with sexual content, rather than blanking or page form editing. One of the main fact observed is the constant change on the pronouns used to refer to transgender scientists and also to represent partners of LGBT+, such as found on Sophie Wilson. This can even lead to a semi-protected block on a page, as notice on Tim Gill's article.

For the most popular pages we notice that insults were more present on Alan Turing's article, whilst on Bill Gate's we have more blanckings and silly edits. Analyzing the insults themselves, we found that both had sexual content accusing them of being homosexual, with the pejorative use of the gay word, regardless of their sexual orientation. This vandalism is a classical example of the homophobic environment found on engineering and computing fields.

IV. Conclusion

Although society has slowing changing to a more opening and acceptant community we could still see that the web environment is still toxic. LGBT+ representation among Computer Scientists on Wikipedia is small and tends to attract prejudice and vandalism. Moreover, popular non LGBT+ scientists articles also reveal this prejudice by suffering vandalism with insults of sexual content with the pejorative use of the gay word. For future work we intend to try to automate the searches, expand to other scientists and make a deeper comparison of quality and quantity of pages. We also intent to interview LBGT+ in computing, in order to understand how this fact has impacted their professional live.

References

[1] G. C. Kane, "It's a network, not an encyclopedia: A social network perspective on wikipedia collaboration." in *Academy of management proceedings*, vol. 2009, no. 1. Academy of Management Briarcliff Manor, NY 10510, 2009, pp. 1–6.

[2] K. Smets, B. Goethals, and B. Verdonk, "Automatic vandalism detection in wikipedia: Towards a machine learning approach," in *AAAI workshop on Wikipedia and artificial intelligence: An Evolving Synergy*, 2008, pp. 43–48.

[3] S. M. Mola-Velasco, "Wikipedia vandalism detection," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 391–396.

[4] P. Shachaf and N. Hara, "Beyond vandalism: Wikipedia trolls," *Journal of Information Science*, vol. 36, no. 3, pp. 357–370, 2010.

[5] N. Hara, P. Shachaf, and K. F. Hew, "Cross-cultural analysis of the wikipedia community," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 10, pp. 2097–2108, 2010.

[6] A. Halfaker, J. Morgan, Y. Pandian, E. Thiry, W. Rand, K. Schuster, A. Million, S. Goggins, and D. Laniado, "Breaking into new data-spaces: Infrastructure for open community science," in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. ACM, 2016, pp. 485–490.

[7] K. Panciera, A. Halfaker, and L. Terveen, "Wikipedians are born, not made: a study of power editors on wikipedia," in *Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 2009, pp. 51–60.

[8] G. Van Rossum and F. L. Drake, *Python library reference*. Centrum voor Wiskunde en Informatica, 1995.

[9] W. McKinney, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.