Hatred Apparatus: A Sculpture for the First Half of the 21st Century

Fabrizio Augusto Poltronieri

Creative AI Lab, Institute of Creative Technologies, De Montfort University
fabrizio.poltronieri@dmu.ac.uk

German Alfonso Nunez

Universidade de São Paulo
gancgana@gmail.com

Nicolau Centola

UNESP
centola.nicolau@gmail.com

Abstract

This paper presents a sculptural object, titled "Hatred Apparatus" (2014–), an artwork that regularly scrapes user comments from news websites and Twitter accounts and classifies them using a neural network according to the intensity of hatred they display. Comments classified with a confidence index higher than 0.6 in the hate speech or offensive language category are entered into a database and randomly displayed every minute on a small black-and-white canvas centred on top of the sculpture. This article describes the creative process, the events that inspired the creation of the device, and the technical issues involved in its creation. It concludes with a reflection on the Hatred Apparatus placement in relation to other works of art.

Keywords

Hate speech, offensive language, sculpture, natural language processing, machine learning.

Introduction

One of the major revolutions introduced with the advent of Internet Technologies (IE) is the possibility for users around the globe to freely communicate at nearly no cost. Information and Communication Technology (ICT) provides inexpensive ways for anyone to instantaneously reach millions of other users. But despite some early optimism, this has led to a substantial increase in hate speech, a term that can be understood as "any kind of communication in speech, writing or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor." (United Nations 2019, 2)

Cohen-Almajor defines hate speech as "bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics. Hate speech expresses discriminatory, intimidating, disapproving, antagonistic and/or prejudicial attitudes towards those characteristics, which include sex, race, religion, ethnicity, color, national origin, disability or sexual orientation. Hate speech is intended to injure, dehumanize, harass, intimidate, debase, degrade, and victimize the targeted groups, and to foment insensitivity and brutality against them" (2011, 1).

One of the main attributes of online hate speech is that it takes advantage of the way many internet websites and virtual communities operate. Anonymity is supposedly an advantage of the Internet as a medium for communication, since individuals are not compelled to reveal aspects of their offline identity unless they wish to (Brown 2018), but ironically it is this very anonymity that allows the proliferation of such hate. Hence, although online anonymity allows opportunities for freer speech, giving voice to people who otherwise would not have it, it also disinhibits users to write things that they would not otherwise state face to face (Suler 2004).

Anonymity and invisibility are key properties that lead to the omnipresent tension between freedom of speech and hate speech. According to Titley, this tension between "understandings of the fundamental importance and scope of 'freedom of speech', and the injustice and implications of hate speech, can never be satisfactorily resolved" (2014, 14).

Although this is a global problem, the ways in which different countries deal with this question are quite diverse. In the United States, for example, hate speech is protected under the free speech provisions of the First Amendment, although its limits have been extensively debated in the legal sphere (Davidson et al. 2017). However, "many liberals in the US tend to object to general hate speech regulations. They believe that legal restrictions on racist or hate speech are not warranted because they violate the speaker's autonomy." (Cohen-Almajor 2018, 39).

In many other countries, however, including the United Kingdom, Canada, Germany, and France, there are laws prohibiting and targeting hate speech. Often people convicted of using hate speech face large fines and even imprisonment. These laws extend to the internet and social media, leading many websites to create their own provisions against hate speech (Davidson et al. 2017). Despite such legislation, many websites still use slow manual moderation to tackle hate speech and offensive language, so abusive comments and posts are usually left online for long periods of time (Gambäck and Sikdar 2017, 85). During the Spring of 2017, parliamentary committees in Germany and the UK strongly criticized leading social media sites such as Facebook, Twitter and YouTube for failing to take sufficient and swift action against hate speech, and the German government threatened to fine these social networks up to €50 million per year if they continue to fail to act (Thomasson 2017).

Origins of the inspiration to build an apparatus that displays only hate speech

The topic of hate speech started to draw our attention in 2013, when massive protests emerged in our home country, Brazil. Various individuals and social groups from across the ideological spectrum took part in the demonstrations. What started as a movement demanding free public transportation soon became an entanglement of multiple demands that exposed cracks in the Brazilian social fabric. These protests, in other words, uncovered tension that Brazilians had tried to hide since at least since the end of military rule in the 1980s.

As Saad-Filho argued, these protests "expressed a wide range of demands about public service provisions and governance, and concerns about corruption. Their social base was broad, starting with students and left-wing activists, and later including many middle-class protesters and specific categories of workers. The deep and contradictory frustrations expressed by the protests were symptomatic of a social malaise associated with neoliberalism, the power of the right-wind media, the limitations of the federal administration, led by the Worker's Party (PT), the rapid growth of expectations in a dynamic country, and the atrophy of traditional forms of social representation." (2013, 657).

In the end, the protests weakened president Dilma Rousseff's position, and paved the way for her impeachment some two years after the beginning of Brazil's social unrest. A short time later, she was replaced by her former ally and vice-president, Michael Temer, a traditional politician, whose deeply unpopular government slashed public services, flouting the wishes of the majority of Brazilians. In sum, this was the background that allowed for the rise of Jair Bolsonaro, an ultra-conservative figure, who was elected president in 2018.

A former army captain turned politician, Bolsonaro branded himself a political outsider, an anti-politician (Arantes 2020), despite having been involved in politics for almost 30 years. His notoriety, as Faley reminds us, "comes from making a series of bizarrely offensive statements during his career" (2019, 4). His long list of offenses includes: telling a fellow legislator that she was too ugly for him to rape her; saying that he would rather have a dead son than accept him as gay; and taunting Afro-Brazilians, indigenous communities and those from the poorer states of the Brazilian northeast. He also stated that the Brazilian dictatorship's only mistake was that it did not kill enough of its political opponents. In his inauguration speech Bolsonaro vowed to "liberate" Brazil from "socialism", "gender ideology", "political correctness" and "ideologies that defend criminals." (Faley

2019). We believe it is this same aggressive and supposedly "carefree" attitude, shared by other extreme right-wing populists around the world, that has fostered an explosion of online hate speech in Brazil's social media.

Hatred Apparatus

When the 2013 protests began, the division between left-wing and right-wing supporters in Brazil came into view both in traditional media and online. One of its most visible signs was seen in the huge influx of hate speech in Brazilian social media and the comment sections of news websites. This fact led us to research hate speech internationally and compare that to what was happening nationally in Brazil. This research was the starting point for "Hatred Apparatus", an artwork we have been developing since 2014 as part of a series of creative AI apparatus, such as "LoveApparatus" and "Prophecy Apparatus".

The basic aim of Hatred Apparatus is simple: to display internet comments identified as not only hateful but also containing extremely offensive language by a machine learning system. The first version was exhibited in the 2017 edition of the renowned artistic festival Ars Electronica, in Linz, Austria.

The apparatus (figure 1) comprises a wood box, featuring a small black-and-white display, which houses a NVIDIA® Jetson Nano™[1], a small, powerful computer designed to run neural network applications. The initial versions of the Hatred Apparatus ran on Raspberry Pi computers. The software consists of a series of custom Python programs, developed mostly by us, that constantly scrape the comments of users and readers from a number of websites and Twitter accounts in English, covering the entire political spectrum, from the extreme left to the extreme right, including Young Communists, News and Letters Committees, Communist Party USA, AlterNet, Common Dreams, Consortium News, The Intercept, Daily Kos, Mint Press News, OpEdNews, Raw Story, World Socialist Web Site, The American Conservative, The American Spectator, the American Thinker, Breitbart, City Journal, Daily Caller, The Daily Wire, Fox News, the Foundation for Economic Education, Free Republic, Hot Air, the National Review, The New American, NewsMax, One America News Network, PowerLine, Quillette, Reason, RebelNews, RedState, and Ricochet.

Web scraping is a technique to automatically access and extract large amounts of information from a website. Our scraping script, coded using the Python libraries *Beautiful Soup*[2] (to scrape websites) and *snscrape*[3] (to scrape Twitter), makes no distinction regarding the collected content. Working alongside the scraping process, another Python program, using machine learning, analyses the comments, and stores those identified as containing either hate speech or offensive language with a confidence higher than 0.6 in a MongoDB database. This process runs twice a day, and it is controlled by a Linux cron job.

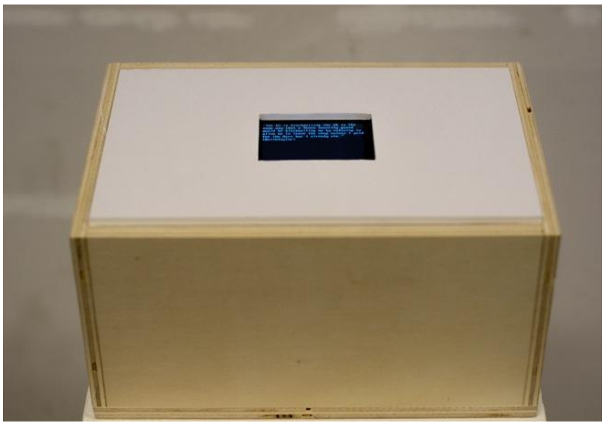

Fig. 1. *Hatred Apparatus*, 2014–2021, +zero, Wood box, LCD screen, NVIDIA® Jetson Nano™.

The data stored in MongoDB is then randomly displayed on the apparatus' screen, and its messages are rotated every minute. Although much more information about every comment is stored in the database, including its source, article title, URL, author, date it was scraped, and type of comment (hate speech or offensive language), only the text of the selected comment is displayed, with no other information or context. Figure 2 depicts the entire Hatred Apparatus workflow.

[1] https://developer.nvidia.com/embedded/jetson-nano-developer-kit

[2] https://www.crummy.com/software/BeautifulSoup/

[3] https://github.com/JustAnotherArchivist/snscrape

As a work of art, Hatred Apparatus acts as a repository and living memory of one of the most harmful aspects of our digitally interconnected global society. The fact that online hate speech is so frequently and easily found means that it is deeply embedded in various sectors and layers of today's societies. The internet just gave these people an anonymous platform to spread their hate and intolerance towards those considered different or somehow, the enemy.

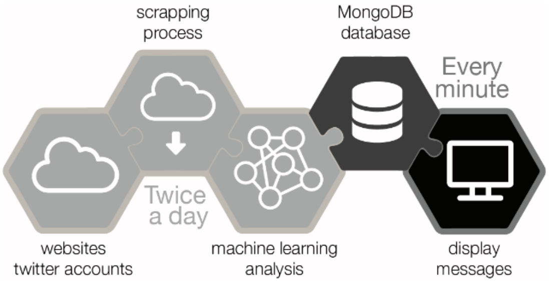

Fig. 2. *Hatred Apparatus' workflow*.

It is important to highlight the fact this artwork is not a celebration of hate speech. It is the very opposite. As artists, it was our deliberate decision to collect and show one of the ugliest facets of human nature.

Automatic detection of hate speech, and abusive and offensive language

The task of identifying online hate speech, and abusive and offensive language has been a central topic in many research communities for more than 20 years (Gambäck and Sidkar 2017). Even with recent advances in the field of machine learning, it is still a great challenge to classify texts and filter out hate speech using machine learning alone.

Machine learning has been shown to solve several language-processing tasks, such as part-of-speech tagging, sentiment analysis, and entity recognition (Gambäck and Sidkar 2017). Therefore, despite the fact that "the majority of the solutions for automated detection of offensive text rely on Natural Language Processing (NLP) [. . .], there is lately a tendency towards employing pure machine learning techniques, like neural networks for that task." (Pitsilis et al. 2018, 2)

Nonetheless, the complexity of the natural language (Badjatiya et al. 2017) and the fact that "what is considered a hate speech message might be influenced by aspects such as the domain of an utterance, its discourse context, as well as context consisting of co-occurring media objects (e.g., images, videos, audios), the exact time of posting and world events at this moment, identity of author and targeted recipient" (Schmidt and Wiegand 2017) further complicates matters. As "a key challenge for automatic hate speech detection […] is the separation of hate speech from other instances of offensive language" (Davidson et al. 2017, 1), we decided to extend the scope of the artwork to also include abusive and offensive language, and not to rely only on strictly defined hate speech. Therefore, our artwork does not require a high precision algorithm, and the inclusion of abusive and offensive language in general made the database and its contents even richer, reflecting the polarities and emotions at play.

Bespoke vs. off-the-shelf solutions

Although there are many datasets annotated for hate speech, online abuse, and offensive language (Vidgen and Derczynski 2020; The Alan Turing Institute 2020; Gilbert et al. 2018), the process of designing, training, and optimizing neural network models is time consuming, and it was not our main focus. Creating a new dataset was also out of the question, as the annotation process is costly in terms of time and resources. To name a few examples that illustrate the complexity of creating a dataset, we can cite Sood et al. (2012), who collected 1.6 million comments from the Yahoo! social news website, 6,500 of which were randomly selected for annotation by 221 people on Amazon Mechanical Turk (AMT); Xiang et al. (2012) created offensive language topic clusters using logistic regression over a set of 860,071 tweets automatically annotated using a boot-strapping technique and supplemented with a dictionary of 339 offensive words; Wasseem (2016) discusses a similar issue while providing a set of 6,909 English hate speech tweets annotated using Appe[4] (former

[4] https://appen.com/

CrowdFlower) users; and Risch et al. (2020) reports that its *toxic comments* dataset contains about 220,000 comments, each labelled with regard to six non-exclusive classes (toxic, severe toxic, insult, thread, obscene, and identity hate).

Over the years, various techniques have been employed by the Hatred Apparatus to classify its scrapped messages. First, we tried our own implementations using word filters. This method, as expected, produced mediocre results, which led us to the implementation of more sophisticated NLP techniques, such as Simple Surface Features using character n-grams, Word Generalization and Sentiment Analysis. These also produced mixed results, resulting in many false positives.

Recently, along with the replacement of the Raspberry Pi by the NVIDIA® Jetson Nano™, message classification has been achieved by Python's *HateSonar*[5] library, which allowed us to easily detect hate speech and offensive language without the need for further training. It also provides a confidence percentage alongside the hate speech and offensive language classifications, which we use to decide which messages should be included in our database, using the 0.6% threshold.

Technically, the adoption of an off-the-shelf solution for the classification of messages has proved correct, since *HateSonar* is a specialized library for this task, and its development was based on one of the most influential research studies on the subject, namely, the pre-trained model based on the work of Davidson et al. (2017).

Conclusion

Our apparatus, in its essence, shows only what is already visible on the surface of social networks and news websites. These are public messages seen by millions on a daily basis. What interests us, however, is the aesthetic exploration of the crude and the deplorable, and how absurd these messages are, especially when observed in isolation, without any context, in a sculptural object that remits nothing to the universe of hate from which these messages emanate. In this regard, our artwork approaches photographs or paintings of war and misfortune, which serve as an artistic and social repository of challenging times, without glorifying the content in any way, and its display of decontextualized utterances points to the absurdity and even stupidity of its content. Since the messages are usually written in a crude and immature manner, the collection produced by the apparatus also highlights the interchangeable nature of these distilled pieces of anger, which could be used in basically any heated debate in social media. These are spectral messages that haunt us daily, collected and exhibited to heighten the effect of reality without any attempt to intervene in their capacity to reflect current events. Here we propose a time machine for future generations, an eyewitness to our incapacity to deal with our current situation. Hence, the act of collecting and displaying these images, we believe, is one that is especially relevant to our current age, where the rise of political extremism and hate speech is undeniable (Bleich 2011; Hunter and Power 2019; Butt and Khalid 2018; Lazaridis *et al.* 2016).

References

Arantes, Marilia. 2020. "A president without a party: Bolsonaro's strategy to depoliticize Brazil," accessed January 6, 2021, https://www.opendemocracy.net/en/democraciaabierta/bolsonaro-president-without-party-strategy-depoliticize-brazil/.

Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta, and Vaseduma Varma. 2017. "Deep learning for hate speech detection in tweets," https://arXiv.org/abs/1706.00188.

Bleich, Erik. 2011. "The rise of hate speech and hate crime laws in liberal democracies." *Journal of Ethnic and Migration Studies*, 37, no. 1: 917–34.

Brown, Alexander. 2018. "What is so special about online (as compared to offline) hate speech?" *Ethnicities* 18, no. 3:297–326.

Butt, Khalid and Momiyar Khalid. 2018. "Rise of far-right groups in Trump's America." *Journal of Political Studies* 25, no. 2: 105–20.

[5] https://github.com/Hironsan/HateSonar

Cohen-Almagor, Raphael. 2011. "Fighting hate and bigotry on the internet." *Policy & Internet* 3, no. 3, Article 6.

Cohen-Almagor, Raphael. 2018. "Taking North American white supremacist groups seriously: The scope and challenge of hate speech on the Internet." *International Journal for Crime, Justice and Social Democracy* 7, no. 2: 38–57.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated hate speech detection and the problem of offensive language," https://arXiv.org/abs/1703.04009.

Faley, Conor. 2019. *In spite of you: Bolsonaro and the new Brazilian resistance*. New York: OR Books.

Gambäck, Björn and Utpal Kumar Sikdar. 2017. "Using convolutional neural networks to classify hate-speech." *Proceedings of the first workshop on abusive language online*, 85–90.

Gilbert, Ona, Naiara Perez, Aitor García Pablos, and Montse Cuadros. 2018. "Hate speech dataset from a white supremacy forum." *Proceedings of the second workshop on abusive language online,*11–20.

Hunter, Wendy and Timothy Power. 2019. "Bolsonaro and Brazil's illiberal backlash." *Journal of Democracy* 30, no.1: 68–82.

Lazaridis, Gabriella, Giovanna Campani, and Annie Benveniste. 2016. *The Rise of the Far Right in Europe*. London: Palgrave Macmilian.

Pitsilis, Georgios, Heri Ramampiaro, and Helge Langseth. 2018. "Detecting offensive language in Tweets using deep learning," https://arXiv.org/abs/1801.04433

Risch, Julian, Robin Ruff, and Ralf Krestel. (2020). "Offensive language detection explained." *Proceedings of the second workshop on trolling, aggression and cyberbullying* pages, 137– 43.

Saad-Filho, Alfredo. 2013. "Mass protests under 'left neoliberalism': Brazil, June-July 2013." *Critical Sociology* 39, no.5 :657–69.

Schmidt, Anna and Michael Wiegand. 2017. "A survey on hate speech detection using natural language processing." *Proceedings of the fifth international workshop on natural language processing for social media* pages, 1–10.

Sood, Sara, Elizabeth Churchill, and Judd Antin. 2012. "Automatic identification of personal insults on social news sites." *Journal of the American Society for Information Science and Technology* 63, no. 2: 270–85.

Suler, John. 2004. "The online disinhibition effect." *Cyber Psychology and Behaviour* 7: 321–326.

The Alan Turing Institute. 2020. *Online hate research hub*, accessed on January 10, 2021, https://www.turing.ac.uk/research/research-programmes/public-policy/online-hate-research-hub.

Titley, Gavan. 2014. "Hate speech online: Considerations for the proposed campaign". In *Three studies about hate speech and ways to address it*, edited by Gavan Titley, Ellie Keen, and László Földi. Strasbourg: Council of Europe.

Thomasson, Emma. 2017. *German cabinet agrees to fine social media over hate speech*. Reuters, April 5, accessed on January 10, 2021, http://uk.reuters.com/article/idUKKBN1771FK.

United Nations. 2019. "United Nations strategy and plan of action on hate speech," accessed on January 6, 2021, https://www.un.org/en/genocide prevention /documents/UN%20 Strategy%20and%20Plan%20of%20Action %20on%20Hate%20Speech%2018%20June %20SYNOPSIS.pdf

Vidgen, Bertie, and Leon Derczynski. 2020. "Directions in abusive language training data: Garbage in, garbage out," https://arxiv.org/abs/2004.01670.

Waseem, Zeerak. 2016. "Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter." *Proceedings of 2016 EMNLP workshop on natural language processing and computational social science*, 138–42.

Xiang, Guang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. "Detecting offensive tweets via topical feature discovery over a large scale Twitter corpus." *Proceedings of the 21st ACM international conference on information and knowledge management,* 1980–984.

Biographies

Fabrizio Augusto Poltronieri (São Paulo, 1976) is an award-winning computer artist, researcher and curator with a special interest in the relationships between Art, Design, Digital Media, and Technology. His expertise lies in the development of creative coding and its exchanges with philosophical questions. Two of his artworks from the "Visual Theogonies" series (Dionysus and Calliope, 2014) are in the V&A's – Victoria and Albert Museum – collection, in London, UK. Poltronieri is a member of the IOCT (Institute of Creative Technologies) and Associate Professor in Creative Technologies at De Montfort University, Leicester, UK. He is currently researching Creativity & Artificial Intelligence, applying machine and deep learning techniques to the production and design of narratives, moving images and objects.

German Alfonso Nunez is currently a FAPESP postdoctoral fellow at the University of São Paulo, Brazil. Previously he was a Visiting Postdoctoral Scholar at Stanford University. Co-editor of the recent Handbook of Popular Culture and Biomedicine: Knowledge in the Life Sciences as Cultural Artefact (Springer, 2018), he is interested in the intersections between specialist technoscientific knowledge, politics and cultural products. His latest research attempts to explain the reach and impact of Cold War American policy into the development of technological/digital art in Brazil during its military period. Alongside his academic work, he is also a member of the computational artist trio known as [+zero], nominated for the Brazilian Contemporary Art PIPA awards of 2011.

Nicolau Centola was born in 1967 in Ribeirão Preto, Brazil. PhD at the Institute of Arts at UNESP, with a thesis on chance in sound art. Master in Education, Art and History of Culture at Universidade Presbiteriana Mackenzie, studying the installation Poème Électronique. Lecturer since 2005 in several colleges in São Paulo, Brazil, Centola develops works in the areas of sound art, computer art, digital art, performances and installations.