DETECT THEN ACT

TECHNICAL REPORT 2

ISSN 2736-6391

CREATIVE COUNTERNARRATIVES AGAINST HATE SPEECH

TOM DE SMEDT
TEXTGAIN
ANTWERP, BELGIUM
TOM@TEXTGAIN.COM

STEF LEMMENS
HOGESCHOOL PXL
HASSELT, BELGIUM
LEMMENS_STEF@HOTMAIL.COM

ADAM COOKE
WREXHAM GLYNDWR UNIVERSITY
WALES, UNITED KINGDOM
ADAM.COOKE@GLYNDWR.AC.UK

GUY DE PAUW
TEXTGAIN
ANTWERP, BELGIUM
GUY@TEXTGAIN.COM

Standing up to online hate speech and bullying requires a lot of confidence and creativity. Responses should not be toxic or target personal traits, yet to have some kind of impact they should be thought-provoking and prickly. We present a number of real-life case studies, and ideas to augment these with CC techniques.

INTRODUCTION

Online social media generate new content like there are raindrops in a storm. Over 10 million pictures appear on Facebook every hour (Internet.org, 2013) and over 20 million messages on Twitter (Twitter.com, 2013). No doubt, this includes a wealth of enjoyable cat pics and cooking recipes, but it has also become clear that part of the content is undesirable, fueling polarization, civil unrest and violent extremism in our societies. To illustrate this, in May 2016 the European Commission agreed with Facebook, Twitter and YouTube on a Code of Conduct for countering illegal hate speech online (European Commission, 2016), and in June 2017 the German government passed a Network Enforcement Act (NetzDG) to counter hate speech and fake news (Deutsche Welle, 2018). These regulations were in part a response to jihadist groups abusing Twitter and YouTube as propaganda machines (Tomé, 2015), and in part an attempt to curb the rise of reactive co-radicalization by far-right extremists (Douglas, 2015).

METHODS

To address the problem, two broad approaches are being used in the field: automatic detection of hate speech by machines, and counternarrative campaigns by humans.

AUTOMATIC DETECTION

Various technological solutions have been proposed, typically using Machine Learning (Schmidt & Wiegand, 2017). However, these also pose ethical and legal challenges regarding fairness, freedom of expression, and privacy. Local legislation as to what is illegal can differ worldwide (Jaki & De Smedt, 2019), and new top-down regulations by governments are sometimes seen as problematic.

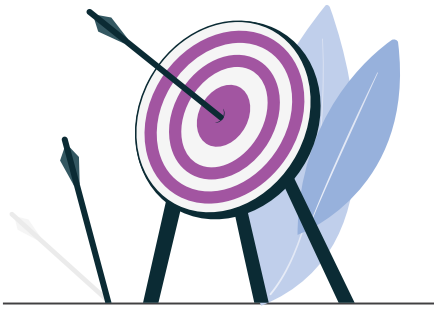
To illustrate this, the Hate Crimes and Hate Speech Bill introduced in 2016 in South Africa has also caused concern over potential abuse to limit unpopular political speech (Human Rights Watch, 2017).

To explain why automatic detection is necessary in the first place, consider the following case study. Using our own system, we compared two sets of 100,000 random Facebook and Twitter messages written in Dutch, the language for which the system is most reliable. The first set contains random messages from 2015, the second from 2020. In these sets, the number of offensive messages discovered rises from about 5,000 in 2015 (5%) to 15,000 in 2020 (15%), with three times more racist expressions (1.5%) than in 2015 (0.5%). Not all of these messages are equally offensive of course. In 2020, we find a subset of 500 extremely offensive messages (0.5% or 1/200). If this reflects the global situation, it would mean that online hate speech is on the rise, and that every hour over a 100,000 extremely offensive messages appear on Twitter (1/200). The main point is that it is no longer possible to process such volumes by hand. Hence, automatic detection tools can be useful for discovery, to make a first selection.

COUNTERNARRATIVE CAMPAIGNS

Many research initiatives also focus on educating media literacy and promoting **self-regulation** among social media users, with a mix of activism, fair reporting, and sometimes AI. One of these is ours, funded by the European Commission. It uses a combination of automatic detection and human activism. The AI is used as a "Sorting Hat" to discover offensive content, but it is then left up to the activists to decide how to deal with it (ignore, respond, report). The general idea is not to coat offenders with tar and feathers, but to counter their opinions with facts and irony. This task is challenging: offenders can post whatever they want, whereas responders cannot.

Typically, offensive messages come in the form of "edgy memes", meaning prejudiced, rude and/or aggressive cartoons. Popular propaganda tactics include **ridicule**, for example by highlighting an unflattering snapshot of a face, reinforcing **stereotypes** (e.g., people of color framed as primitives), **aggression** (booting Muslims) and **symbolism** (folkloristic images from The Good Old Days, etc.).



FIGURATIVE LANGUAGE TO THE RESCUE!

It seems as if human creativity is endlessly wicked when it comes to demeaning fellow human beings. The **shock value** of this kind of content can be quite effective (Dahl, Frankenberger & Manchanda, 2003), to the extent that it is difficult to suppress a surprised laugh at the cruelty, sarcasm and hatred on display. Essentially, from the comfort of an anonymous internet connection (Colleoni, Rozza & Arvidsson, 2014), the effect is not so different from watching a stranger trip over a banana peel and chuckling at his discomfort. But in our project responders cannot riposte in kind with content that ridicules, aggravates, or dehumanizes, which often takes the sting out of the response. "That is very mean of you" is not likely to leave an impact. In other words, our unsolved challenge is how to generate thousands of responses (counternarratives) that are creative, witty and thought-provoking, without being offensive.

Perhaps some lessons can be learned from history. For example, in the past writers have resorted to allegory, metaphor, incongruity, and poetry to get their message across in trying times (cf. Orwell's *Animal Farm*, Golding's *Lord of the Flies*). Such techniques are well-understood by the CC community (cf. Veale, Shutova & Klebanov, 2016).

RESULTS

The remainder of this paper presents three real-life case studies from our project (there are more) and the approaches that we have tried so far. Most of our approaches constitute curated ontologies and manual craftsmanship, with a potential for being automated by CC researchers.

CASE STUDY 1: CC PSEUDONYMIZATION

Activists in the project have a dashboard at their disposal that displays recent toxic messages, along with a Respond button. This dashboard uses a **double-blind** approach. The identity of offenders is kept hidden from the activists to avoid prejudice, and vice-versa responses are posted on social media from anonymous accounts that guarantee the privacy of the activists. But we also want to track who the **repeat offenders** are, since usually only about 5% of the offenders are responsible for polluting an entire debate (Kreißel, Ebner, Urban & Guhl, 2018). Such accounts are assigned a memorable pseudonym like *Cathy-hates-cats* or *Rudy-likes-to-rumble*. The problem is that a large number of pseudonyms is needed to avoid collisions (i.e., two real names being assigned the same pseudonym), whereas a large search space may also generate inappropriate pseudonyms like *Toby-touches-toddlers*.

A related problem is the production of anonymized avatars in the dashboard (see **Figure 1**). These avatars are generated by taking a picture of a face, mirroring it so that the face will face the viewer, and pixelating the result. This approach may be prone to de-anonymization with today's AI. Ideally, an infinite amount of truly anonymous pixel faces is needed to avoid collisions. One solution might be to use the "best" avatars as training material for a Generative Adversarial Network (GAN) to produce variations.

Figure 1. Example anonymized avatars.

[1] https://artbreeder.com

CASE STUDY 2: CC COUNTERNARRATIVES

Our second and arguably most pressing challenge is to assist activists in coming up with witty responses to thousands of offensive messages. So far, we have developed a curated ontology of handcrafted responses (e.g., "You must be fun at parties", "We don't have time to play with you now"), along with a prototype generator for incongruity ("This post is as exciting as a hot air balloon at an aerobat ics show"). Without doubt, there are much more elaborate approaches beyond our expertise that can be successful.

A team of art students is also designing memes. **Figure 2** shows an example of our Demonic Cat. This image was produced by blending a cat pic with a picture of an octopus in Artbreeder, an online GAN.[1] It can be understood as a metaphor for how unpleasant the targeted offensive message is. Since the art students are constrained by what variations they can produce (i.e., their artworks should be non-offensive) we noticed that several students resorted to depicting animals instead of real people, using accompanying wordplay ("Are you kitten me?"). We can also see a range of CC opportunities here.

Figure 2. Example countermeme (Artbreeder).

CASE STUDY 3: CC POETRY

Perhaps the ethically most interesting approach is to serve the offenders their own content. It is elegant in the sense that no additional prejudice is introduced. One artist affiliated with the project maps toxic messages into poems "by stripping the right words". Volunteers are asked to submit offensive messages to the artist, who then erases certain words to create new sentences, to "manipulate hate speech into poetry that retains the emotion of the original author, but in a more suggestive and less short-sighted way than before, and in doing so question its meaning".

[1] https://artbreeder.com

Figure 3 illustrates the creative process. First, the words that carry the most meaning are identified. These words are usually nouns and verbs. Then, preceding or succeeding words can be included to form a syntactically correct sentence. In some cases, no sentence can be formed. The artist remarks that: "It is important for me to keep the original elements intact, meaning that I will avoid adding letters or words to make the new sentence more readable. But sometimes you are missing just one letter or word to create the perfect poem and you have to try different combinations, or otherwise discard the entire message. I do allow myself to find new words within existing words, which makes it possible to change the narrative of the message in a powerful way". In the given example, the original message reads: "C'mon you just kill the guy, lift him out his bed at night, chain him, throw him in the trunk, drive to the woods in Germany and leave him there naked." The transformed poem can be read as: "Simply broken, just stuck, shed the darkness there on display".

Figure 3. Example hate speech poetry.

The artist reports rewarding experiences with using this creative process to teach media literacy to school students, showing them how a message can become negative or positive through different word combinations. Some school students introduce their own techniques by erasing or adding letters, allowing them to experiment with language use.

An idea for automatic hate speech poetry

Transforming a hateful message into a pensive poem could be a fascinating CC challenge. Here perhaps is one possible approach:

- Discover *n* toxic messages, using available AI.
- Generate all possible permutations of *n* messages, or *m* candidate poems.
- Rank the *m* candidates using Neural Language Modeling, e.g., Dutch RobBERT.[2] This will rank candidates by semantic and grammatical coherence and provide log-probabilities for each message.
- Ask human reviewers to select the *k*-best candidates, where *k* is dependent on the number of human reviewers available.
- Post the poem and observe reactions, retweets, likes, etc., to learn threshold *x*.
- Candidates with a log-probability higher than *x* can be trusted more and sent out with less human review.

[2] https://github.com/iPieter/RobBERT

An idea for automated multimodal layout

Can we produce infinite variations of multimodal memes? The idea of automated layout is not new. Karl Gerstner's seminal work Designing Programmes (1963) already outlined a number of approaches informed by computational developments of the time. For example, his *Morphological Typography* proposed a system for categorizing expressive typographic characteristics, using descriptive terms such as appearance (e.g., size, proportion), color (shade, value) and technique (spacing, form). The parameters could be automated to produce generative variations. Gerstner expanded on these concepts throughout the book, reflecting on photography, layout, grid systems, and so on. Nowadays, these ideas are well-understood in the Generative Art community, which has close ties to the CC community.

The art students in our project (who are not part of the GA community) initially engaged themselves by mapping how toxic memes can be classified *visually*, to understand their construction and their ability to be appropriated for viral variants. It is clear that the relation between text and background image is vital to that understanding. The background image is usually appropriated from popular culture or from the history of art, with infinite possibilities for reappropriation by adding "the right" text (see Figure 4).

Variations in meaning can be generated through rapid iterative developments of either the text or the image. Our students' try-outs can offer a roadmap for new CC bots:

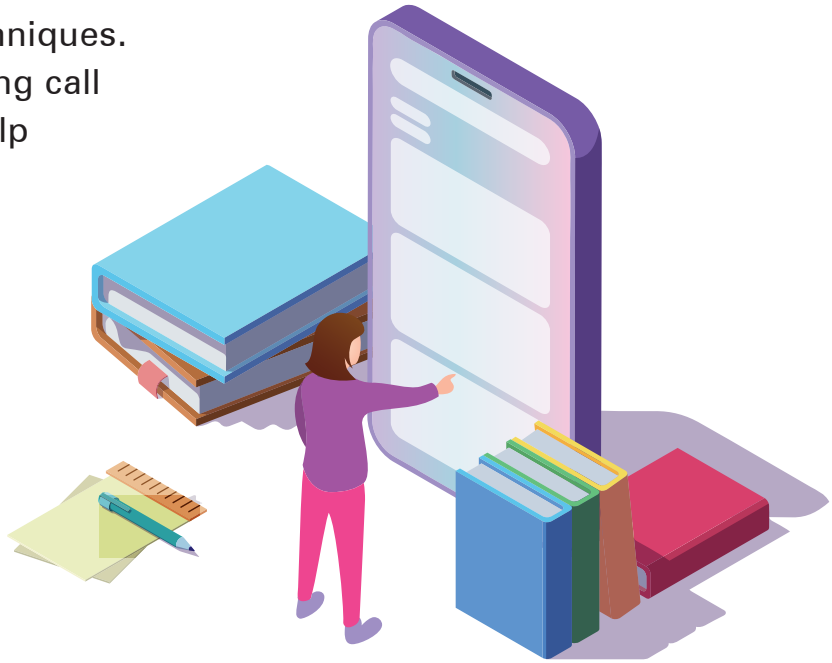
Figure 4. Example automated multimodal layout.

- Scale output to meet demand: reply to waves of propaganda with computationally less expensive designs.
- Standardize and automate countermeme creation. This will require an analysis of available tools: Processing for example is not deployable on Facebook and Twitter, yet precompiled GIFs are.
- Build ethical evaluation into the bot, constraining core elements (e.g., kittens = yes, pitbulls = no).

Various meme-making websites provide images where text can be overlayed at speed. No previous design experience is necessary, a series of simple steps are conveyed to allow users to come to a visual conclusion based on a limited set of predetermined parameters. Generally, this is how toxic memes are bred, and there is room for improvement when producing countermemes. A performative task which could be automated to allow intelligent counternarrative responses to offensive images posted online.

DISCUSSION

In this paper, we have attempted to give a brief overview of community-based efforts for tackling offensive content on online social media, where we can see a central role for operationalizing existing CC techniques. Our paper is meant as an inspiring call for CC researchers that could help us to tackle the challenge.



ACKNOWLEDGEMENTS

This project is co-funded by the Rights, Equality and Citizenship Programme of the European Union (2019-2021). The content of this report represents the views of the author only and is his/her sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

REFERENCES

Colleoni, E., Rozza, A., and Arvidsson, A. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication, 64*(2), 317-332.

Dahl, D. W., Frankenberger, K. D., and Manchanda, R. V. 2003. Does it pay to shock? Reactions to shocking and nonshocking advertising content among university students. *Journal of advertising research, 43*(3), 268-280.

Deutsche Welle. 2018. Germany implements new internet hate speech crackdown. https://www.dw.com/en/germany-implements-new-internet-hate-speech-crackdown/a-41991590

European Commission. 2016. The EU Code of conduct on countering illegal hate speech online. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Human Rights Watch. 2017. South African Move on Hate Speech a Step Too Far. https://www.hrw.org/news/2017/02/21/south-african-move-hate-speech-step-too-far

Internet.org. 2013. A Focus on Efficiency: A whitepaper from Facebook, Ericsson and Qualcomm. http://citeseerx.ist.psu.edu/viewdoc/download?-doi=10.1.1.695.46 02&rep=rep1&type=pdf

Jaki, S., and De Smedt, T. 2019. Right-wing German hate speech on Twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518.

Kreißel, P., Ebner, J., Urban, A., & Guhl, J. 2018. Hass auf Knopfdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz. Technical Report, Institute for Strategic Dialogue, London. https://www.isdglobal.org/wp-content/uploads/2018/07/ISD_Ich_Bin_Hier_2.pdf

Pratt, D. 2015. Islamophobia as reactive co-radicalization. *Islam and Christian–Muslim Relations 26*(2): 205-218.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media,* 1-10.

Tomé, A. 2015. The Islamic State: trajectory and reach a year after its self proclamation as a Caliphate. *e-journal of International Relations, 6*(1).

Twitter. 2013. New Tweets per second record, and how! https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html

Veale, T., Shutova, E., and Klebanov, B. B. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies, 9*(1), 1-160.