WARNING: PLEASE SKIP THIS ARTICLE IF YOU'RE EASILY OFFENDED BY PROFANITY

SWEARING IN CONTEXT

"NEW HOPE FOR THE WORLD. GOD BLESS AMERICA AND OUR HOMES. HAVE NO SWEARING, BOYCOTT PROFANITY! PLEASE DO NOT SWEAR, NOR USE OBSCENE OR PROFANE LANGUAGE. THESE CARDS ARE FOR DISTRIBUTION. SEND FOR SOME - THEY ARE FREE."

185 E 76 st New York, N.Y.

If you were given a pink card with the message to the left a hundred years ago, it meant the Anti-Profanity League has been alerted to your vulgar language. This pious and impractical band, founded in 1901 by Arthur Samuel Colborne of 185 East Seventy-sixth, distributed these cards to further their goal of totally eradicating all swearing – the very epitome of pissing in the wind.[1]

Nowadays, you are unlikely to receive a printed card, but your social media posts may be flagged if they contain offensive or hateful speech. The scale and instant reach of social networks mean that moderating language is no longer the job of a handful of eccentrics, but the task of the latest AI technology for flagging harmful content, and tens of thousands of contract workers across the globe.

For sure, unlike with the Anti-Profanity League, moderating Facebook, Twitter and YouTube are not motivated by linguistic squeamishness: online speech can have real-world consequences. However, the Anti-Profanity League exemplifies the pitfalls of an broad-brush approach to recognising harmful content.

Firstly, what should be censored? Colborne took an expansive approach to this, wanting to prohibit not only swear words, but also 'leaders-on', such as hell, devil take it, dad burn it, gee whiz, and doggone, for fear such words would act as gateway drugs to the truly filthy ones.

However, as we pointed out in the previous instalment of our profanities series, swearing is not always intended to offend, and the same swearword can have multiple meanings.

"They found examples of tweets where this was used to verbally abuse another user (you are an ass), to emphasise a feeling (a good ass day) and express emotion (pain in the ass). It was also used as an auxiliary (really need someone to save my ass), as a marker of identity (now this is a group of ass-kickers) and in a non-vulgar way, given the context (Kick-Ass 2 – what a movie)."

We used our profanity filter (a list of common swearwords) to extract several thousand profanity-containing responses from customers with faulty internet service. We then applied our best judgment as to whether we should continue to communicate with this customer, pause the conversation for a while, or stop the conversation altogether. We found that for 30% of responses we should just continue as usual. In a less contentious and more social setting than talking to your internet service provider, a far greater proportion of swearwords would be indicative of something other than abusive intent.

So, what was in this 30% that triggered the profanity filter? A proportion were milder swearwords, perhaps the equivalents of Colborne's 'leaders-on' such as crap or bloody. But the most obscene words can be repurposed:

'The tech was fucking brilliant' one happy customer opined.

[1]https://stronglang.wordpress.com/2017/02/26/joseph-mitchell-a-s-colborne-and-the-anti-profanity-league/
[2]https://www.engadget.com/2017/11/07/twitter-lgbt-search-block-explanation/
[3]https://www.theverge.com/2018/6/4/17424472/youtube-lgbt-demonetization-ads-algorithm
[4]https://www.theguardian.com/world/2018/jul/05/facebook-declaration-of-independence-hate-speech
[5]https://www.wsj.com/articles/how-to-use-irony-on-the-internet-11565409660
[6]https://www.aclweb.org/anthology/C16-1231.pdf
[7]https://www.aclweb.org/anthology/N19-1221.pdf

These examples demonstrate some of the challenges facing any AI algorithm to identify offensive comments. An algorithm that over-indexes on the presence of particular words ignores the nuances of how language is used. Twitter ran into this problem when it inadvertently censored tweets containing the words ***bisexual*** or ***gay***, because its algorithm mistook such words as indicating adult-content[2]; YouTube has also been accused of demonetising LGBT content based on particular words such as ***trans***[3]. Algorithms also struggle to recognise the difference between words used in quotation versus in anger: last year Facebook's algorithm censored the Declaration of Independence due to a passage describing indigenous Americans as "merciless Indian savages."[4]

The issue here is that the meaning of a sentence is not just a function of its words. It's informed also by context, which can come in many forms: the intention of the author, background knowledge, the surrounding text, and the intended audience.

We need algorithms that can incorporate the full history of user's posts, social connections, and a wealth of background and historical knowledge. This, unsurprisingly, is extremely difficult. AI researchers are working on incorporating contextual information into its algorithms by directly learning profiles of authors of text. This authorship information has already been shown to benefit sarcasm detection, a similar task to profanity detection. Like swearing, sarcasm is often an example of what internet linguist Gretchen McCulloch calls a 'trust fall' – a linguistic test used to engender trust between interlocuters akin to falling backwards trusting your friend to catch you.[5] In a 2016 paper[6], researchers showed that including context of a user's previous tweets could dramatically improve the ability of an AI model to judge whether a future tweet was sarcastic or not. A Facebook paper[7] on content moderation shown at NAACL (an NLP conference) this year also uses author features to improve model accuracy – in this case the features are learnt using an effect called homophily: the tendency for similar people to be connected in a social network. Using a technique called graph convolutions, user representations can be learnt both from a user's posts and the online communities in which he or she interacts, helping the model better differentiate between, say, homophobic abuse and gay self-expression. This approach, however, comes with its own problem: the risk of finding people guilty merely by their association with others.

Euan Matthews
Director of AI and Innovation
Contact Engine
contactengine.com

Knowing where context is relevant (and when not) is beyond current AI, meaning human moderators are going to be greatly needed for a long time. Currently, just 16% of bullying and harassment posts are proactively detected by Facebook's technology before they are reported by users. Of course, moderating the posts of 2 billion Facebook users is much harder than customer responses, but the following lesson still applies: we need to use a combination of technology and human empathy to understand how best to treat customers – particularly when they start swearing.