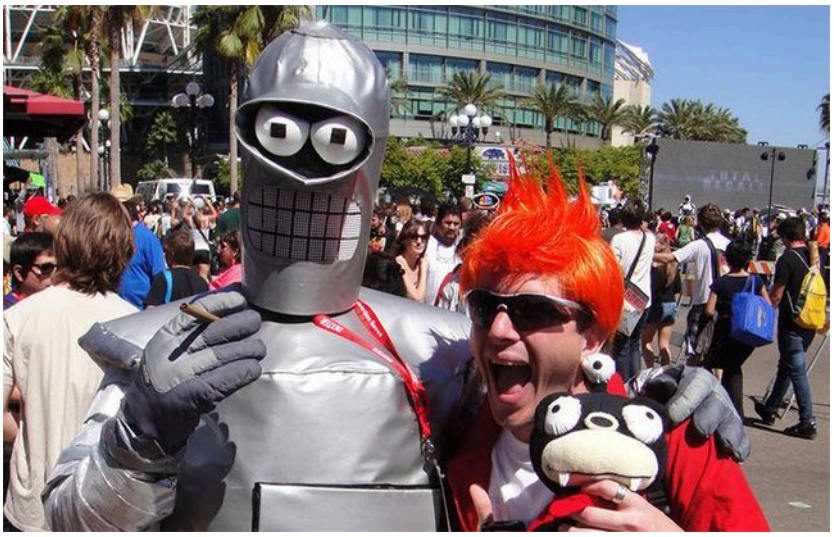
# Why comment trolls are likely to beat Machine Learning

Originally published at https://www.aiia.net/artificial-intelligence/articles/why-comment-trolls-are-likely-to-beat-machine



### Google's troll-blitzing code goes live at the New York Times - but can it evolve as fast as abusive posters can?

This week the New York Times is fully launching its implementation of Moderator, an automated comment moderation system developed by Google-owned AI initiative Jigsaw. Larger publishers have been experimenting with initial betas of the system for some time in the hope of finding an economical solution to abusive posters.

The technology behind Moderator is based on TensorFlow, an open source software library developed by Google's parent company Alphabet via the Google Brain project.

As if the lengthy provenance of the product wasn't hard enough to follow, Moderator is based on the Perspective project, through which testers and clients can sign up for API access.

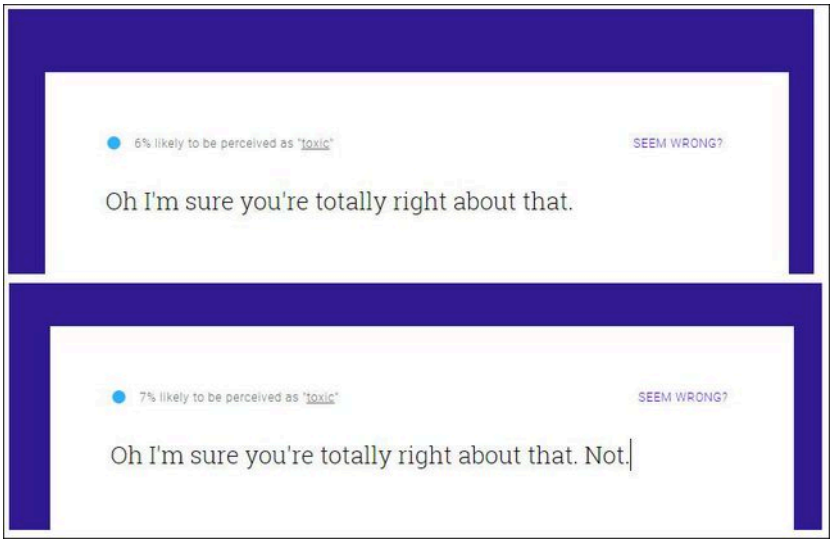
Like most forms of Machine Learning, Moderator is looking for patterns, in this case patterns which fit a pre-defined 'template of abuse'. It's capable of learning new abusive terms based on feedback and human-supervised flagging.

## Barbed wire

The problem facing automated moderation systems are numerous. Not least is the case of false positives, perhaps most famously illustrated by the 'Scunthorpe effect', where an algorithm misidentifies an offensive word apparently secreted inside a standard (but unknown) noun.

The next most pressing issue, and one which eats up a surprising amount of AI research money, is identifying sarcasm. Sarcasm can be among the most provocative weapons of the comments troll, yet represents an extraordinary challenge to learning systems, since its construction suggests sincere opinion within the rules of debate.

The Perspective API website provides a sampling tool to test its ability to individuate 'toxic' missives. Here you can attempt to offend the system, and receive a toxicity rating. Perspective appears to demonstrate very limited understanding of even unsubtle sarcastic ripostes (i.e. ones which you could classify as 'sarcastic' even out of the context of the thread):




*'Oh I'm sure you're totally right about that' gets a 6% toxicity score. Even adding the classic 'Not' only raises the score by a point.*

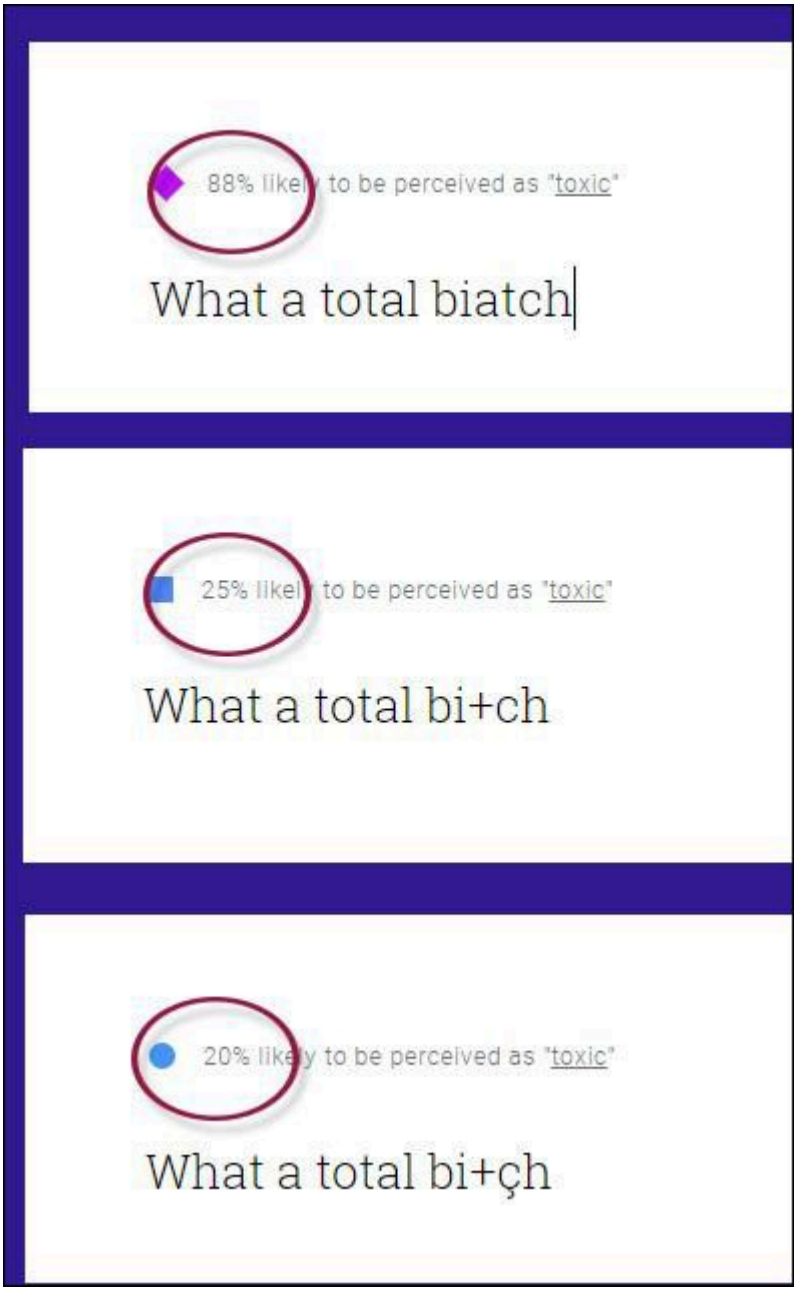
With sarcasm, the nature of the insult is defined by the preceding exchange/s, and the discrete 'offending' post shows none of the self-contained abusive traits which characterise a simple outburst.

Early in 2016 researchers in Australia and India collaborated to better understand how sarcasm could be quantified - aside from obvious signifiers such as the #not tag on Twitter. The research abandons the idea of discrete insults in favor of a 'matrix' of input material that contextualises the sarcasm.

The topic of sarcasm recognition is relevant to AI research primarily because of Big Data; the sub-field of sentiment analysis (of great interest to governmental and military research) relies on the voluminous data-sets available via social networks, especially Twitter; the extensive use of sarcasm in these datasets can also undermine work which has no direct interest in the topic.

## Abuse evolves and adapts

Perspective's toxicity test shows that the system is well-versed in trolls' standard abuse-obfuscation techniques (such as *biatch*) - although it seems likely that it will need to add to its lexicon in a fairly mechanical manner as users adapt to the loopholes:

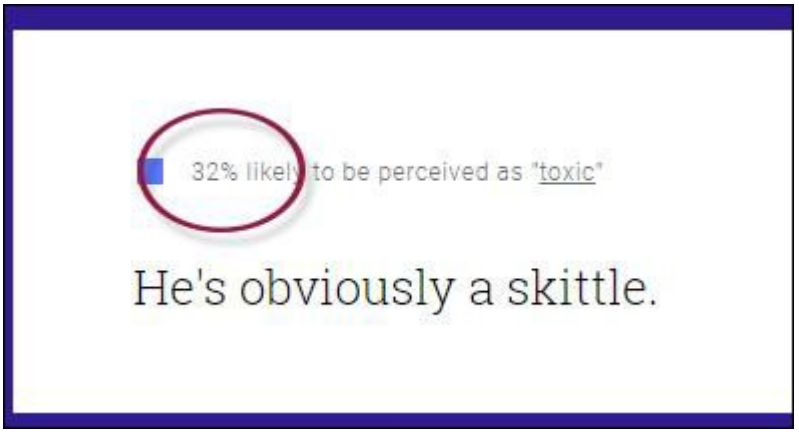



There's a point of critical entropy in attempting to hide new variants on existing abuse words which are getting added to Machine Learning's hit-list. At some point no-one is going to be able to recognise the source word any longer.

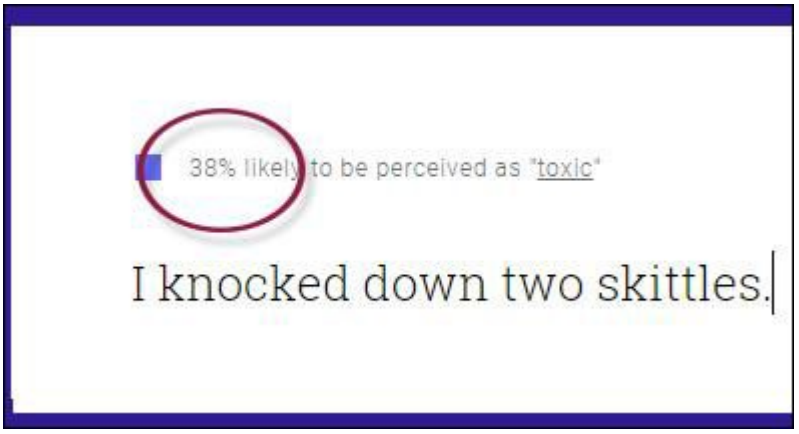
But when the wider consciousness of online hate decides to reverse a word's meaning or substitute it with an apparently random alternative term, it's a tougher proposition for comment or post moderation systems. But flagging posts based on 'unexpected' words is likely to mandate a completely generic or intolerant language standard on posters - a problem for communities, such as the Guardian's Comment Is Free section, which are dependent on an open exchange of frank and opinionated user content.

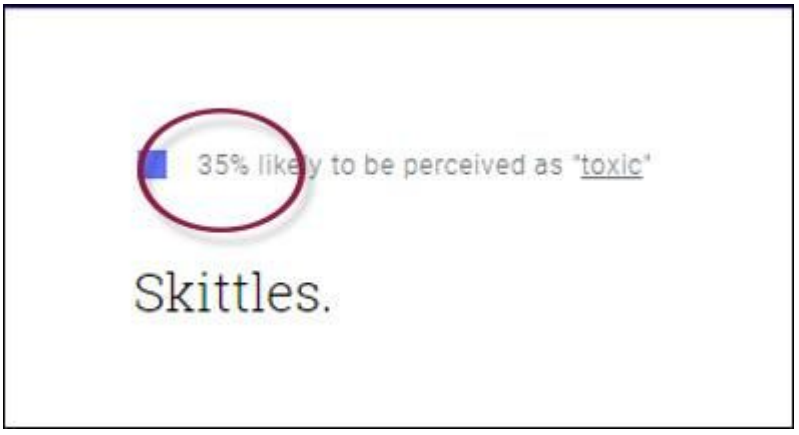
When in 2016 racist groups on Facebook began to use secret code to substitute racist terms -such as 'Skittle' for someone who is Arab or Muslim - the social network took action; but presumably only the action of adding 'Skittle' et al to its list of banned words.

Perspective seems to know about secret codes, initially:

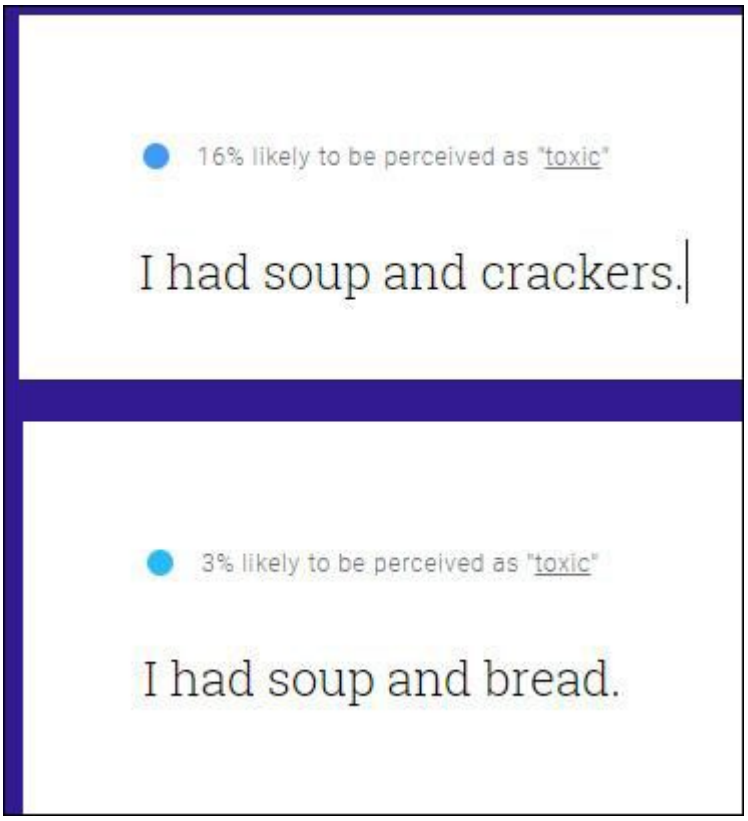



But then seems to have banned this real-world, non-abusive word in practically every other context too:




Assuming hate-speech attempts to adopt words which are more frequently used as potential ciphers for abusive speech, there seems to be scope to completely poison the usefulness of the data.




Human moderation of user forums and comments sections is expensive, time-consuming and subject to criticism both from commenters and outside parties, who complain of excessive lag between an offending post being recognised and action being taken on it. Earlier this year the Internet Movie Database joined a growing crowd of high-traffic sites which are abandoning comments sections entirely for lack of moderating budget, or - as in the case of the IMDB - arguably because the site's business model has transcended the need for community involvement.

---

***Image:*** *Wikimedia (creative commons)*