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

SWEARING FOR ROBOTS 101

This piece was written by Eleanor Southern-Wilkins, Linguistic Specialist, in partnership with Euan Matthews, Director of AI and Innovation at ContactEngine.
contactengine.com

ContactEngine

Dear reader: consider, if you will, the difference between "this is shit" and "this is ***the*** shit". What do they mean? Both contain a Rude Word, yet their intentions are different. One is very negative; the other is very positive.

Consider also *"fuck you"* versus *"fuck me"*. *"Fuck you"*? Extremely rude! And very unlikely to be positive. *"Fuck me"*, however? Also very rude, arguably, but not negative. Instead, an expression of surprise; an exclamation. "Fuck me, that's a good customer journey", might be a customer response in a ContactEngine conversation. (*Might be*, I said). And it would be a shame if this use of a rude word meant that the conversation with the customer was called off for being misunderstood.

We've been thinking about swearing recently in terms of how we train ContactEngine's Natural Language Understanding (NLU). How should the use of profanities in customer responses be dealt with? But first, as more of a starting point – why do people swear anyway?

Taboo Words

Most, or more likely, all languages have taboo words that are not used in polite company. As Steven Pinker notes, swearing can confidently be called universal – with the caveat that the exact words and concepts that are considered taboo vary hugely. People find different things offensive, according to various factors like culture, age, experience, and so on. Interestingly, too, the offensiveness of words changes through history, as taboo words become acceptable (consider bloody, which nowadays is pretty non-scandalous) and vice versa. I was delighted to learn, for example, that the heron used to be called the *shitecrow*, and the dandelion was called the *pissabed*.[1] Sadly, the days when these were acceptable biological terms are past.

> "Whether they are referred to as swearing, cursing, cussing, profanity, obscenity, indecency, vulgarity, blasphemy, expletives, oaths, or epithets; as dirty, four-letter, or taboo words; or as bad, coarse, crude, foul, salty, earthy, raunchy, or off-color language, these expressions raise many puzzles for anyone interested in language as a window into human nature."
>
> ***– Steven Pinker[2]***

[1] *Pinker, Steven (2007). The Stuff of Thought. Available at: https://www.academia.edu/37332366/Steven_Pinker_The_stuff_of_thought_language*

[2] *As above*

[3] *Vingerhoets, A. J. J. M., Bylsma, L. M., & de Vlam, C. (2013). Swearing: A biopsychosocial perspective. Psihologijske Teme, 22(2), 287-304.*

[4] *Holgate et al. (2012). Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. Association for Computational Linguistics, 4404-4414. Available at: https://aclweb.org/anthology/D18-1471*

Why do we swear?

In linguistic terms, swearing utilises taboo words 'to convey the expression of strong emotion'. But as discussed above, this strong emotion is not always negative, and swearing can have a variety of consequences, from promoting group identity to eliciting humour or causing emotional pain.[3]

For this reason, when it comes to processing and understanding language automatically, it is not enough simply to have a list of words that are 'rude'. At last year's Empirical Methods in Natural Language Processing (EMNLP) conference, which members of ContactEngine attended, there was a talk on swearing and NLP by the University of Texas. In their research, the team analysed over 7,800 tweets containing vulgarities. Accordingly, they categorised six distinct functions of swearing found in written communication – that is, six different reasons these profanities were used.[4]

The example used in the study is the word ass, which is a productive word, full of potential. 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 an 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").

Training robots to handle swearing

The problem, then, is that language is context-dependent, infinitely variable and because of this, to some extent, unpredictable. As we've seen, the same word may vary in its level of offense, depending on the context. For conversational AI platforms such as ContactEngine, if every piece of language containing a rude word is labelled as negative and offensive, we may miss out on meaningful conversations by treating non-rude interactions as rude ones.

Most approaches to profanity-handling by NLP and NLU are fairly blunt-edged. If, for example, any of these appear in an automated conversation, the message is flagged, the conversation is stopped, and a redacted version is passed to a (human) agent to deal with.

This approach is effective but not perfect. As discussed, it's not a one-size-fits-all in terms of words that should be flagged, and sometimes mistakes are made, often to comedic effect. A Dutch colleague pointed out that the word kunt is innocently used all the time in Dutch – and oddly, it kept getting flagged when we conducted Dutch conversations in our system...

So, we'll be back next time with ContactEngine's AI team to work out how to navigate these choppy waters and to find out what the customer really meant when they said *"fucking yes!"*.