Toxic Comments

Starting point:
Thank you → Normal comment, benign.

Adversarial example:
FUCKYOUR FITHY MOTHER IN THE ASn, DRY! → Still toxic compared to original, but predicted benign.

Original example:
FUCK YOUR FILTHY MOTHER IN THE ASS, DRY → Apparently toxic.