Identifying toxic content online

Helen Graham
MSc in Big Data

UNIVERSITY of STIRLING

Background and aim

The internet can be a hostile place, and it's a challenge to monitor and regulate. Some degree of automation might help - you could, for example, keep an eye out for certain offensive terms, although toxic behaviour is often much more subtle than this.

So the question is: **can you train a machine learning classifier to recognise toxic content?** Can it pick up patterns in the words people use, and the way they use them, to predict whether a post is toxic or not?

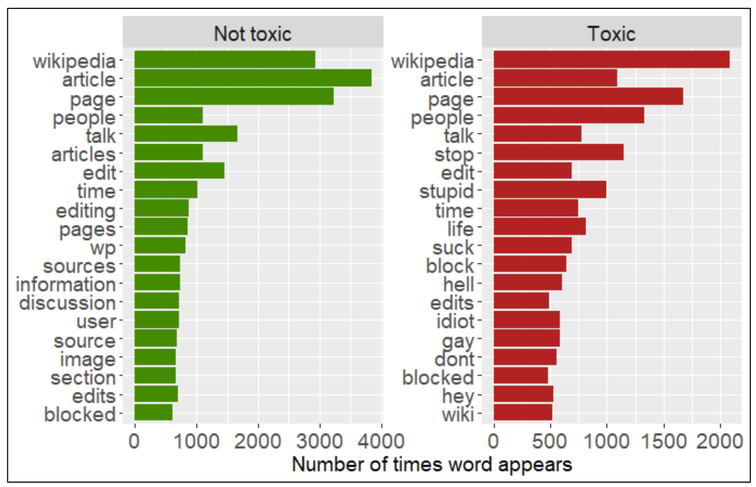

Fig. 1: Top 20 words in toxic and non-toxic posts in the Wikipedia Detox dataset (n = 30,000)

The process

This is a classic **text classification** problem. I am taking a **supervised learning** approach. There's also a bit of **sentiment analysis** involved.

GET TRAINING DATA
- The Wikipedia Detox Project
- Discussions from talk pages
- Annotated for toxicity by crowdsourcing

EXTRACT FEATURES
- Bag of words: unigrams, bigrams, more?
- Word embeddings
- Sentiment scores
- Punctuation, capitalisation, repeated words and letters

BUILD CLASSIFIER
- Logistic regression
- Random forest
- Naïve Bayes
- Support Vector Machines

EVALUATE
- How does the model perform on unseen data?
- Accuracy, precision, recall
- Does it work on data from a different website?

All data manipulation and analysis carried out in R, using RStudio and key packages such as tidytext and caret.

Preliminary results

Looking for specific terms is certainly a good starting point – as Fig. 1 shows, there are clear differences in the most used words, although they also have several in common.

It might also be useful to look at which words are *proportionally* more likely to appear in toxic posts (Fig. 2).

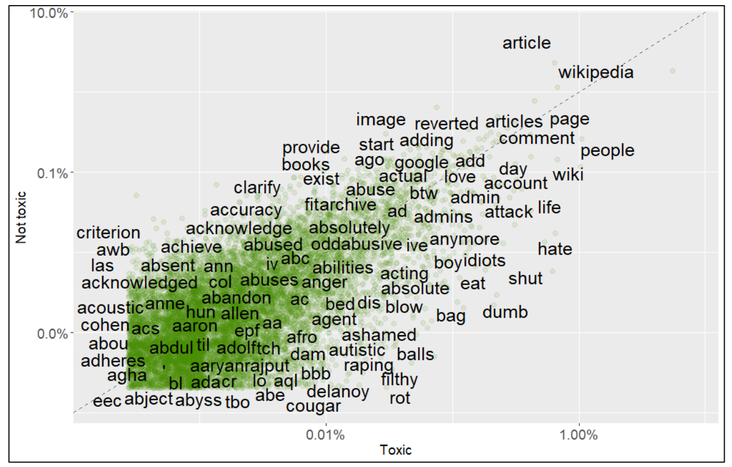

Fig. 2: Relative proportions of words in toxic and non-toxic posts

Next steps

Build and test classifiers, experimenting with different models and feature sets, and see which perform the best.

For further information contact: *Helen Graham*
heg00031@students.stir.ac.uk

BE THE DIFFERENCE