ABUSIVE CONTENT DETECTION

# Objective

To detect when a comment from a conversation would be considered insulting to another participant in the conversation. Sample used are texts drawn from conversation streams like news commenting sites, magazine comments, message boards, blogs, text messages, etc. The idea is to create a generalizable single-class classifier which could operate in a near real-time mode, scrubbing the filth of the internet away in one pass.

Perform text analytics on the comments and classify them into abusive and not abusive.

# Description

Sample of around 6000 comments is taken for analysis and validation. Sample contains chat messages of different users from various sources. These samples are cleaned and used for statistical analysis to build an algorithm that could identify abusive comments.

Sample chats are cleaned by removing stop words and punctuations. Sentences are tokenized to individual words using Natural Language Processing packages in Python. The individual words are tokenized and stemmed (converted to their basic form). Term Frequency Inverse document frequency matrix is used to assign numerical values to the cleaned/treated words. Relevant features are extracted from this matrix using Chi square test. These features are used to train a logistic regression model. Supervised learning is performed, and the model is then used to classify 2200 new chat comments.

# Result

Model successfully classified the new chat comments into abusive and not abusive category with 81% correctness.

Cases of inaccuracy majorly include typos that changes the meaning of sentence, short chats with abusive words that are not meant to abuse the reader, use of words that convey multiple meanings and abuse to third person.

More than 50% of the comments that included abusive words are not actually abusive. Abusive word variants are used in many cases instead of direct abusive words.