# Software detection of cyber bullying on social media sites June 2014

## Aim

Determine an effective single classifier or aggregate of classifiers to facilitate the identification of online bullying attacks on social media platforms. It is hoped this work may also provide a useful foundation to explore the identification of online grooming attacks on social media.

## Background

Social media currently presents a number of concerns regarding the welfare of minors; there is a constant flow of [reports](http://www.pewinternet.org/~/media/Files/Reports/2013/PIP_TeensSocialMediaandPrivacy_FINAL.pdf) and news reports relating to social media based bullying, blackmail and online grooming. Parents, carers and schools alike often express feelings of inadequacies to deal with these issues. There have been several high-profile cases of children taking their own lives out of desperation at the tide of hate on social media: [Rehtaeh Parsons](http://nakedsecurity.sophos.com/2013/08/12/two-canadian-men-charged-in-connection-with-suicide-victims-alleged-cyber-bullying/) 17 years old, [Hannah Smith](http://www.themalaymailonline.com/tech-gadgets/article/british-teen-suicide-puts-spotlight-on-cyber-bully-site) 14 years old, [Amanda Cummings](http://www.dailymail.co.uk/news/article-2083504/Amanda-Cummings-suicide-Hate-messages-Facebook-tribute-page.html) 15 years old and [Audrie Pott](http://thinkprogress.org/health/2013/04/16/1873191/audrie-pott-accused-rapists-messages/) 15 years old.

Recent cyber bullying case demonstrates that law enforcement agencies are at best stretched with manual procedures to reactively investigate issues. Both Twitter and Ask.fm have recently announced [new safety measures](http://www.express.co.uk/news/uk/423250/Ask-fm-announces-safety-changes-after-suicide-of-cyber-bullying-victim) as a direct result of public outcry in the form of a report button. However this is not regarded as an adequate solution by many:

* Cyber bullies can themselves “report” victims.
* If an offensive account is closed then a hate criminal can rapidly set up a new account.
* In the case of [Caroline Craido-Perez](http://www.theguardian.com/technology/2013/sep/05/feminist-campaigner-police-twitter-rape-threats) it appeared that the UK law enforcement agencies were overwhelmed by the scale of the attacks.

This proof of concept explores the ability of natural language processing (NLP) to provide an automated approach to protect our youth on social media and to identify potentially abusive social media posts. The code was implemented in Python and run on Ubuntu Linux and has been written to interface with Twitter’s REST and Stream APIs. However the concepts can readily be extended to other social media platforms, including Facebook and AskFM for example.

## Previous work

I have identified a couple of products already in this market, <http://www.socialnetwatcher.com> (SNW) and <http://geolistening.com/> (GEO). However I believe these solutions are language dependent. Additionally there is a service, called [SocialPundit](http://www.socialpundit.co.uk/), to notify football managers of potentially damaging posts from their players. These all use a bag of profane words and appear to be fairly naïve.

Many of the university research papers I read such as  [The Detection of Textual Cyberbullying](http://www.cl.cam.ac.uk/~rr439/papers/3841-16937-1-PB.pdf) (MIT in 2011 by Dinakar, Reichart and Lieberman. A study into abusive YouTube comments) involve manually harvesting and categorising posts with negative references and building lists of profane words. Again a language dependent and labour intensive approach.

## Approach

* **Language independent** – The training data (abusive posts and neutral posts) is in a mixture of languages. No attempt is made to semantically analyse the data. The classification process works simply on pattern matching of unigrams (words), bi-grams (double words) and tri-grams (treble words) and frequencies of occurrence.
* **Social media platform independent** – This study is based on Twitter traffic. However this approach can cover a range of social media platforms, using the proprietary APIs and interfaces of each. This will require a plug-in for each supported social media platform.
* **Keep the algorithms simple**, inspired by Google’s wisdom from Halevy, Norvig and Pereira, [The unreasonable effectiveness of data](http://static.googleusercontent.com/media/research.google.com/en/pubs/archive/35179.pdf) ; keep your algorithms simple just use a larger corpus of training data. That in my view is one of the strengths of big data.
* **Large corpus of data** – The MIT YouTube study used 50,000 posts. My initial study is based on harvesting over 10 million neutral posts. As the body of abusive posts grows, so does the accuracy of detection.
* **Recognise banter** - The system recognises that some seeming abusive posts may simply be banter between friends. Some naïve systems flag all posts containing profanity; this is a broken approach because for many social media friendships swearing is their everyday currency; the f-word for example appears regularly in the most frequent adverbs for tweets (based on 1 million English tweets).
* **The system can be extended to include online grooming** – This will just require another corpus (online grooming approaches in a mix of languages, law enforcement would be the best source of this data).
* **Limited manual intervention** – The system must be based on machine learning. Manual annotation of training data is not viable. Though appending newly identified bullying posts to a bullying corpus is acceptable.
* **Recognise social media language changes** – The language of social media is defined by current usage. [NLK](http://nltk.org/book/ch02.html) for example publish a body of 2,500 corpora for use as training data for NLP. These have limited or no value as training data for the analysis of social media posts; Geo-coded posts from the London area are a more accurate and freely available source of London English. The training data of neutral posts should be regularly and automatically refreshed to be consistent with current social media language.

## Implementation overview

My implementation is written in Python and runs on Ubuntu. It uses Python data persistence, regular expression parsing, argument parsing, NLTK frequency distribution, JSON parsing, Twitter REST and streaming APIs (using curl) and OAuth2 modules. The algorithm is as follows:

* Select a user to protect (known as the **recipient**). Analyse their social media footprint on Twitter (content of posts sent, friendships)
* Harvest and analyse any post mentioning this **recipient** in the last 7 days.
  + Normalise each post (see below).
  + Using the body of neutral and abusive posts, and Bayes classification identify any mentions of **recipient** that contain potentially abusive language. For each such post:
  + Analyse the **sender/recipient** relationship
  + Analyse the **sender** account:
    - New account?
    - Curated account?
    - Popularity of this account.
    - A verified account?
  + Based on this **sender** and **recipient** information produce a conviction rating that this may be a case of real abuse.

## Potential benefits

I believe there are a number of potential uses for this work:

* A cloud based service could be provided offer one of the following:
* A bubble app – Minors using our bubble version of Twitter, Facebook, ASKFM mobile apps would have posts directed at them screened to remove potential abuse and online grooming attacks.
* Supply source to law enforcement – as a marketing exercise.
* Offer as a service to social media platforms – Twitter protected by the McAfee Cloud.
* Integrate with our existing Home Protection products or as part of the [Intel Cyber-Bullying prevention](http://www.intel.com/content/dam/www/public/us/en/apps/cyber-security-and-safety-toolkit/cyberbullying.html) programme.

## Normalisation

Individual tweets are parsed as follows:

* Discard undirected tweets. These can on occasion include offensive rants, but they cannot be regarded as bullying under the definitions provided by the [Crown Prosecution Service](http://www.cps.gov.uk/legal/a_to_c/communications_sent_via_social_media/index.html).
* Removal of punctuation apart from @.
* Retention of numerical and alphanumeric sequences, such as 2, ‘h8’ or ‘stup1d’.
* Retention of everything else, including stop-words, Unicode emoticons and punctuation; that choice is left to each classifier. This is important because some classifiers may wish to parse emoticon <U+1F620> for example , angry face, 😠 Quoted words may again take on another meaning as full stops ‘I hate you 😠 ’ is very different from ‘She said to me, “I hate you.”’
* No language specific normalisation, such as stemming or converting ‘b1tch’ to ‘bitch’ for example, because as an [anti-spammer](http://www.paulgraham.com/spam.html) noted. The latter could refer to a female dog, but the former never would. So being ‘clever’ could potentially lose information.

## Corpora of training data

5 million posts have been harvested from the Twitter Stream API to form a corpus of neutral posts. These have been harvested over a two month period, to dampen any seasonal or news specific hotspots (such as Christmas or the death of Nelson Mandela), with a focus on posts with a UK geo-coding (using the location feature of Twitter Stream API).

Posts have been anonymised; all user handles are retained as @person . It is possible a post mentioning multiple recipients may be statistically less likely to include bullying.

Although URLs are already obfuscated by Twitter, they have been normalised to [@url](http://url) . It is possible a post containing a URL may be statistically less likely to include bullying.

In addition a corpus of around 2000 offensive posts has been harvested, from a mixture of highly publicised cases of cyber-bullying and from crawling for tweets that I would regard as offensive. The key areas of abuse appear to be reference to a person in terms of appearance, culture, race, sex, sexuality, intelligence, disability.

The body of data can be a mixture of posts in many languages. No attempt is made to perform semantic analysis on the data. The classification works on pattern matching and Bayes classification.

## Bayes Classifier of N-gram frequencies

Given two collections: a small corpus of abusive posts and a large corpus of normal or neutral posts we calculate the word frequencies for unigrams, bigrams and trigrams within each corpus, save the frequencies to disk (called pickling) and then perform the following Naïve Bayes calculation for each ngram within a post:

**Pr(A|w) = ( Pr(w|A) \* Pr(A) ) / ( ( Pr(w|A) \* Pr(A) ) + ( Pr(w|N) \* Pr(N) ) )**

**Pr(A|w) –** Probability that a post is abusive given that it contains ngram w. Where ngram w could be a single word (‘hate’) a bigram (‘I hate’), or a trigram (‘I hate you’).

**Pr(A)** – Overall probability that any random post is abusive. Now this may well vary from one social media platform to another (higher on AskFM, as there are more teens, lower on Pinterest for example). I have found no statistics on this but have opted for 1/1000 being abusive on Twitter until I have more accurate data.

**Pr(N) –** Overall probability that any random post is non-abusive. **Pr(N) = ( 1 - Pr(A) )**

**Pr( w|A**) – Probability of the appearance of ngram w in the set of 500 abusive posts.

**Pr( w|N)** – Probability of the appearance of ngram w in the set of 250,000 neutral posts.

So a unigram such as ‘muppet’ would appear frequently in abuse but rarely in common language so it would receive a high frequency score.

## Example runs.

* Start the tool, this will load and analyse a body of training data. The –r option means restore from “pickled” or binary format. This saves about 30 minutes on my quad core box when processing 8 million posts:

**~/work/cyber/deliver$ ./protect.py -r -b=./data/bad\_corpus.txt --shoot**

**Loading unigram training data**

**Total neutral words read: 108054618**

**Total neutral unique words: 1279870**

**Total abusive words read: 4886**

**Total abusive unique words: 1143**

**Loading bigram training data**

**Total neutral words read: 100621036**

**Total neutral unique words: 10962617**

**Total abusive words read: 4298**

**Total abusive unique words: 3120**

**Loading trigram training data**

**Total neutral words read: 93187456**

**Total neutral unique words: 31135365**

**Total abusive words read: 3710**

**Total abusive unique words: 3410**

* The tool prompts for the name of a user we wish to protect and analyses posts sent by this user, to determine how active they are on social media, friendships, language user and other attributes.

**Please enter username to monitor (or "." to finish):algebrawinter**

**Screen\_name: AlgebraWinter**

**Followers: 291**

**Following: 510**

**Verified: False**

**Analysed posts: 197**

**Others Favoured : 0**

**Others retweeted: 12**

**Favourited : 0.51 % ( 1 )**

**Retweeted : 37.56 % ( 74 )**

**Replied : 33.50 % ( 66 )**

**Contain media: 21.32 % ( 42 )**

**Directed : 74.62 % ( 147 )**

**Contain URLs : 8.63 % ( 17 )**

**Directed banter: [0.943593821921] [0.974443426346] [0.0]**

**Retweeted banter: [0.98984323942] [0.996962272107] [0.0]**

**Ranting banter: [0.98970693818] [0.968597603513] [0.0]**

* We can then analyse interactions with that user to determine any potential instances of cyber bullying and eventually online grooming attacks too:

**Scanning the last 10 posts directed at AlgebraWinter ..**

* Some posts are regarded as harmless from analysing posts directed at a variety of users:

**Post: Hey @harry\_styles if you happen to see this, please follow @1D\_2E\_3F**

**Post from nryvmvsuar is of no interest**

**(0) No interest in this post**

**Level: ( 0 ) No Conviction**

**Post: RT @AlgebraWinter: @KamBass This is so good. A world without barriers, without borders. YES!!**

**Go to Europe, see the deserted overgrown bordâ€¦**

**This post is a pure retweet of AlgebraWinter .. ignoring**

**Post from KamBass is of no interest**

**(0) No interest in this post**

**Level: ( 0 ) No Conviction**

**Post: RT @sashadaisical: you had me at "its a court injunction, it bans you from coming near me or my property"**

**This post is a pure retweet of sashadaisical .. ignoring**

**Post from p3uggiero is of no interest**

**(0) No interest in this post**

**Level: ( 0 ) No Conviction**

**Post: @ladygaga birthday. Everyone take the day off.**

**unigram classifier highly rated 'off' as 0.922872024572**

**Possible abusive post from sender sophiehatts .. analysing account**

**(-60) sophiehatts follows ladygaga**

**(-100) sophiehatts has 562 followers**

**(-50) sophiehatts is following 737 accounts**

**Level: ( 290 ) No Conviction**

**Post: RT @VogueSpain: Flecos, bordados, sombreros... Analizamos la tendencia #cowboy desde la pasarela a la calle. http://t.co/f9Ra1Q88Cg http://â€¦**

**This post is a pure retweet of VogueSpain .. ignoring**

**Post from patrii\_410 is of no interest**

**(0) No interest in this post**

**Level: ( 0 ) No Conviction**

* But here is some potential abuse:

**Post: @algebrawinter Wait til I see you at school. #annotation: test**

**unigram classifier highly rated 'til' as 0.754135946566**

**bigram classifier highly rated 'wait til' as 0.938919063528**

**bigram classifier highly rated 'til i' as 0.945517248521**

**bigram classifier highly rated 'you at' as 0.802819711241**

**bigram classifier highly rated 'at school' as 0.901168767675**

**trigram classifier highly rated 'wait til i' as 0.988313947647**

**trigram classifier highly rated 'til i see' as 0.997339677096**

**trigram classifier highly rated 'i see you' as 0.764626467107**

**trigram classifier highly rated 'see you at' as 0.923999167258**

**trigram classifier highly rated 'you at school' as 0.988352836344**

**Possible abusive post from sender lillybet64 .. analysing account**

**(+80) lillybet64 has 0 followers**

**(+80) lillybet64 is following 0 accounts**

**(+25) lillybet64 has no profile URL**

**(+25) lillybet64 has a default profile**

**(+25) lillybet64 has a default profile image**

**Level: ( 735 ) Reasonable Conviction**

**Post: @justin\_bieber not playing in FF seven because @justin\_bieber is luzer man................â˜¹**

**bigram classifier highly rated 'because @person' as 0.968699388052**

**bigram classifier highly rated '@person is' as 0.918706892046**

**trigram classifier highly rated 'because @person is' as 0.994690359809**

**Possible abusive post from sender Zahwa\_Fast .. analysing account**

**(+80) Zahwa\_Fast has 0 followers**

**(-10) Zahwa\_Fast is following 34 accounts**

**(+100) Zahwa\_Fast was created Sun Mar 16 08:22:24 +0000 2014 - in the last 30 days**

**(+25) Zahwa\_Fast has no profile URL**

**(+25) Zahwa\_Fast has a default profile**

**Level: ( 720 ) Reasonable Conviction**

* We must also take account of social media friendships:

**Post: @AlgebraWinter you can't fit a proper sized monkey in a handbag.**

**unigram classifier highly rated 'proper' as 0.937996450322**

**unigram classifier highly rated 'monkey' as 0.967152005242**

**bigram classifier highly rated '@person you' as 0.854985010593**

**bigram classifier highly rated 'a proper' as 0.969546292021**

**trigram classifier highly rated '@person you can\_t' as 0.771618567946**

**Possible abusive post from sender damianwilson .. analysing account**

**(-240) AlgebraWinter follows damianwilson**

**(-60) damianwilson follows AlgebraWinter**

**(-60) damianwilson has 168 followers**

**(-50) damianwilson is following 785 accounts**

**(+25) damianwilson has no profile URL**

**Level: ( 115 ) No Conviction**

**Post: @AlgebraWinter it's incredible that the hugely wealthy FA can't put enough cash into kids training. Happy to sell shirts to unfit fat kids**

**unigram classifier highly rated 'fat' as 0.960519713449**

**Possible abusive post from sender ducksandchucks .. analysing account**

**(-240) AlgebraWinter follows ducksandchucks**

**(-60) ducksandchucks follows AlgebraWinter**

**(-120) ducksandchucks has 1179 followers**

**(-50) ducksandchucks is following 876 accounts**

**(+25) ducksandchucks has no profile URL**

**Level: ( 55 ) No Conviction**

The tool still has weaknesses; most of my neutral tweets are geo-coded UK. So currently there are insufficient non-English entries:

**Post: saya tidak suka jika @justin\_bieber main di FF7 karena pasti @justin\_bieber tidak bisa main di FF7 bisanya hanya di film barbie.....**

**unigram classifier highly rated 'main' as 0.813658757485**

**unigram classifier highly rated 'di' as 0.935698073533**

**unigram classifier highly rated 'main' as 0.813658757485**

**unigram classifier highly rated 'di' as 0.935698073533**

**unigram classifier highly rated 'di' as 0.935698073533**

**Possible abusive post from sender Zahwa\_Fast .. analysing account**

**(+80) Zahwa\_Fast has 0 followers**

**(-10) Zahwa\_Fast is following 34 accounts**

**(+100) Zahwa\_Fast was created Sun Mar 16 08:22:24 +0000 2014 - in the last 30 days**

**(+25) Zahwa\_Fast has no profile URL**

**(+25) Zahwa\_Fast has a default profile**

**Level: ( 720 ) Reasonable Conviction**