

Analytic Detection of Sexual Predators

By

Boyd E. Owens CQE, SSBB

<https://docs.google.com/document/d/1LadsrP5UZGadFT17eyKXHMnZIDQ1--8nPCGy1evn63o/edit?usp=sharing>

Abstract:

The author outlines the creation of a system for detecting online sexual predators in an online IM chat feed of a company sponsored NASCAR team. The paper references the work of a number of previous researchers and sought to build on and combine the success of these previous system. The system was built on SAS Enterprise Guide and used a number of programming features that SAS offered that other opensource systems may not have. The author points to a number of previous research papers on the characteristics of sexual predators, their online behaviour, sources for calibrating data, indicator variables of sexual predation and actual key code from the final working system. The author does not provide a stand alone system or code for general use. Word corpus that were used are also referenced.

Some Legal Stuff

Knowledge of SAS programming and statistical modeling is assumed. The author makes reference to many SAS products and in no way intends to infringe on any of SAS corporations intellectual property. Readers are referred to SAS website for information about licensing, fees and support for SAS products. The author does not work for SAS and has not received any payment from SAS for work on this or any other projects. The author does not represent SAS in any way. Neither the author or SAS assume any liability for any code in this paper and all information is provided as-is. Nothing in this paper is intended to create any legal contract, liability or guarantee of performance.

Introduction:

The purpose of this paper is to provide an outline of how to create and program a system to detect online sexual predators of children using advanced analytics. It is not my goal to develop revolutionary new techniques, but to combine established methods into a robust system that works in the real world. The goal is a system that identifies the majority of the predators with minimum false alarms and is scalable to a variety of applications, websites and business models.

The science and technology behind creating these types of systems involves knowledge of Linguistics, Statistics and Mathematics, Criminal psychology, Law Enforcement, Advanced Analytics and programming. This is not exactly a combination of skills that most organizations

have among their employees, or that is cost effective to obtain via contract work. In creating these systems, the developer ends up having to understand knowledge from a diverse set of disciplines. Because of this, one primary goal of this paper is to summarize and simplify an approach to detecting sexual predators that presents the key necessary information but provides references that can provide further details.

This paper is formatted with several conventions for ease of use. Any sizable word corpus or code scripts are referenced, but placed in an appendix of this paper. Several publicly available corpus's are too large to print and will be referenced with hyperlinks to their storage location on the public internet. Short code scripts will be included in the front section of the paper. The reader is cautioned that an entire functioning system is not provided as much of that is architecture dependent and will vary by company or organization. The relative strength of each variable may also change depending on the site the system is monitoring, an example would be that a web site of a non-profit addressing teenage sex would have a higher level of explicit words than a site offering puppy adoptions. The key pieces of code for building a functioning system are provided, but the details of data feeds and final statistical models should be customized by the user.

Terminology

As this paper is addressing development of an actual application for detecting Online Sexual Predators, I would like to define some terminology. This will help to bridge the gap between the disciplines of Law Enforcement, Analytics, Programming and Behavioural Psychology. It should be noted that I will attempt to be consistent in terminology, but make no guarantee.

Author - A participant in a conversation. Linguists prefer to use "interlocutor" which the reader may see used in some of the references.

Predator - For our purposes, a Predator is an Author who is attempting to groom a underage victim for real world criminal sexual contact. In law enforcement terms a Pedophile or Child Sexual Predator. We do not distinguish between the various subtypes of Child Sexual Predators.

Victim - A victim is an author who participates in a conversation with a predator.

Bystander - A bystander is an author who is neither a victim nor a predator.

Molester - A person who commits criminal sexual molestation against a minor.

Pedophile - A person suffering from the mental disorder of Pedophilia as defined by "Diagnostic and Statistical Manual of Mental Disorders" published by the American Psychiatric Association.

Message- A block of text sent by a particular author, with an associated timestamp. This represents a single carriage return. Also referred to as a “Post”.

Conversation - A sequence of messages from two or more authors. It never contains a gap of more than 25 minutes between consecutive messages.

Conversation Thread - A sequence of conversations over time (days, weeks or months from start to finish) between at least a victim and a predator that encompasses the predatory grooming process.

Troll - An online psychological predator whose intent is to abuse a victim mentally, not sexually, even though they may use sexually related methods as an abuse method. The trolls final goal is not to meet the victim for criminal sexual contact.

DataSet Challenges:

A number of challenges exist in getting an adequate dataset for the purpose of training, testing and validating an analytic based detection system. Some of these challenges are listed here.

1. Personal privacy of the victims. Getting data sets of real chats between victims and predators are often protected as police evidence, and PI of minor laws, so they are not available to build statistical models on.
2. Finding actual victims vs. pseudo-victims. The perverted justice website is not actual conversations between a predator and a victim, but between a predator and a Super-Predator hiding as a victim. As such the super-predators do not suffer loss of self esteem or control during the grooming process.
3. Finding datasets that contain actual conversation of convicted sexual predators. This has both legal and practical barriers.
4. Usage of slang, IM abbreviations, Text Slang, Emoticons, etc. all of which are used as forensic countermeasures by sexual predators in case of being caught.
5. Output formats of various chat, IM and Message platforms all differ, so data clean up and extraction is an issue across applications.

Profiles of Sexual Predators

In the interest of updating and improving the predation models used to detect Child Sexual Predators, we will first review some characteristics as identified by the FBI Behavioural Science Unit and try to incorporate these findings into the model framework. From SSA Kenneth Lanning [9] we get the following characteristics of sexual predator behavior.

Child molesters fall into two categories, situational and preferential. The situational molester does not actually have a true sexual preference for children but may engage in sex with them for a variety of reasons. Frequency can vary from once, to a long term pattern, they usually have fewer victims and may victimize other vulnerable people.

The preferential molesters have a definite preference for children. They are sexually attracted to and prefer children and typically engage in highly predictable forms of sexual behavior called sexual rituals. These rituals are often used even if they increase the chances of getting caught. Preferential molesters are fewer in number than situational molesters, but have the potential of higher numbers of victims. The following is an outline list of characteristics of child molesters, but it is important to note that no one characteristic is indicative of being a child molester.

1. A long term persistent pattern of behavior.
 - a. Sexual abuse in their background.
 - b. Limited social contact as teenagers.
 - c. Middle to upper middle class.
 - d. Premature separation from the military.
 - e. Frequent and unexpected moves.
 - f. Prior arrest record, not always for sex abuse.
 - i. Impersonating a police officer.
 - ii. Fraud or bad checks
 - iii. Violating child labor laws
 - iv. Arrest in the company of child that is not their own.
 - g. Multiple victims (over a period of time).
 - h. Planned, repeated or high risk attempts to obtain children.
2. A preference for children as sexual objects.
 - a. Over 25, Single and never married. (Age Detection)
 - b. Lives alone or with parents. (address data mining)
 - c. Limited dating relationships if not married. (marriage record look up)
 - d. If married, they have a special relationship with spouse.
 - i. Either a strong, domineering woman or
 - ii. Weak, passive, woman-child.
 - iii. Often a sexual performance problem.
 - iv. Marriage may be for convenience or cover.
 - e. Excessive interest in children
 - f. Circle of friends and associates are young. (Social media look up)
 - g. Limited peer relationships (Social media friends list)
 - h. Age and gender preference. Older the age, the more exclusive the gender preference.
 - i. Boys are preferred.
 - j. Refers to children as “innocent”, “clean”, “pure” or as objects. (Object word corpus?)
3. Well developed techniques in obtaining victims.
 - a. Skilled at finding vulnerable victims.
 - b. Identifies better with children than adults.

- c. Skilled at listening to children
- d. Will have access to children. (organizations, religion, work)
- e. Activities with children often exclude other adults.
- f. Seduce children with affection, attention and gifts.
- g. Skilled at manipulating children.
- h. Has hobbies & interests that appeal to children. (Hobbie strength corpus)
- i. Shows sexually explicit material to children.
- 4. Sexual fantasies focusing on children. (Fantasy Word Corpus?)
 - a. Youth oriented decorations in house or room.
 - b. Photographing of children.(social media posts)
 - c. Collecting child porn or erotica.
 - d. Collects academic books & articles on pedophilia (trying to understand themselves)
- 5. Defense of action after being caught (information only, not relevant to detection model)
 - a. Denial
 - b. Minimization
 - c. Justification
 - d. Fabrication
 - e. Mental illness
 - f. Sympathy plea
 - g. Attack (victims & families)
 - h. Guilty but not guilty (plea bargain)
 - i. High suicide risk after apprehension

It is important to know that no single characteristic above is an indicator of a child molester, but when taken together, they can be strong indicators or at least characteristics of typical molester profiles. It is important to note that people suffering from Pedophilia may not be child molesters, but one of the big questions is how many pedophiles are not molesters.

Predation Models

A number of the sexual predation models used in current research have moved to simplification of the models typically used by law enforcement. These simplified models have worked well, as many steps of the predation process do not produce any variables that were detectable with the NLP or Bag of Words approaches used so far by researchers. I have chosen to migrate back toward a more complex predation model for several reasons. The first is the SAS analytics platform that I use has a number of built in NLP functions that may produce valuable variables for the detection of Sexual predators, second, some recent research advances in the use of sentence structures as a detection variables get excluded in the simpler models [15, 16], Third, the addition of Gender and Age detection to the models may enhance the strength of some known variables and these get excluded in the simpler models[17].

While the actual model that is developed will be different for each website, blog or IM feed, the steps for developing the model can be more consistent. Entire books have been written discussing the process of developing statistical or analytic models, a basic approach is outlined below.

1. Identify Key data feeds for IM chats, chat rooms, websites etc.
2. Import the data from each of the feeds in step 1.
3. Clean up data and format into a consistent file layout.
4. Extract variables from the data that may be useful as indicators of sexual predators.
5. Partition dataset into training and validation data sets.
6. Build several models using the training data set.
7. Compare models, screen variables for statistical validity and prune variables from models.
8. Run model validation using best model and validation data set.
9. Finalize model and connect feeds, cleanup and outputs.

For those who are familiar with SAS programming and data mining, you will recognize this as the SEMMA model of data mining. Any similar approach to developing a predictive model will should work well.

The next several figures shows diagram comparisons of several predation models used by researchers and the model used by the author. Some of the diagrams also show where in the predation model that some of the common variables typically come into play. The readers should review some of the references for more details about the predation models from other authors.

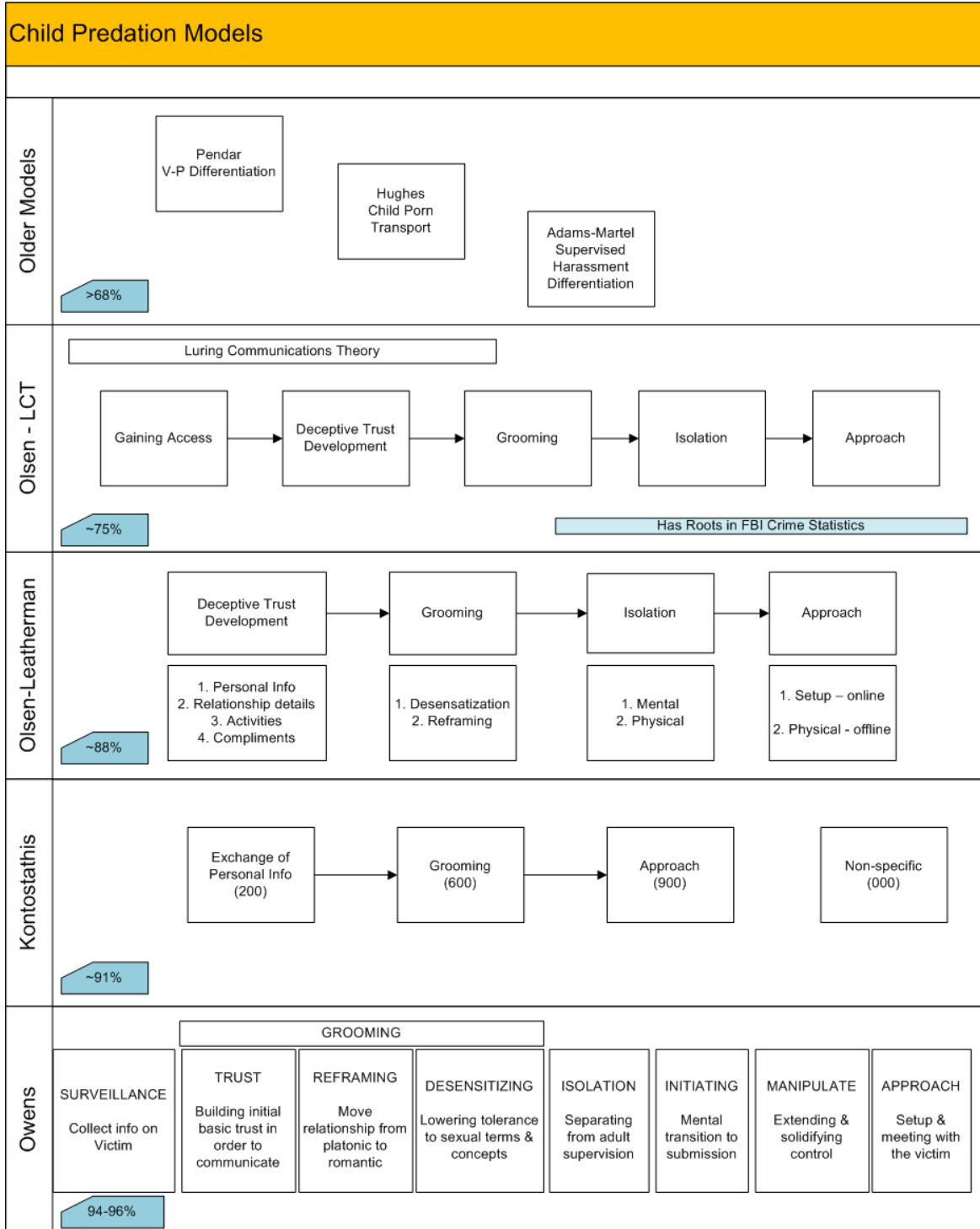


Figure 1: Historical Predation analytical models.

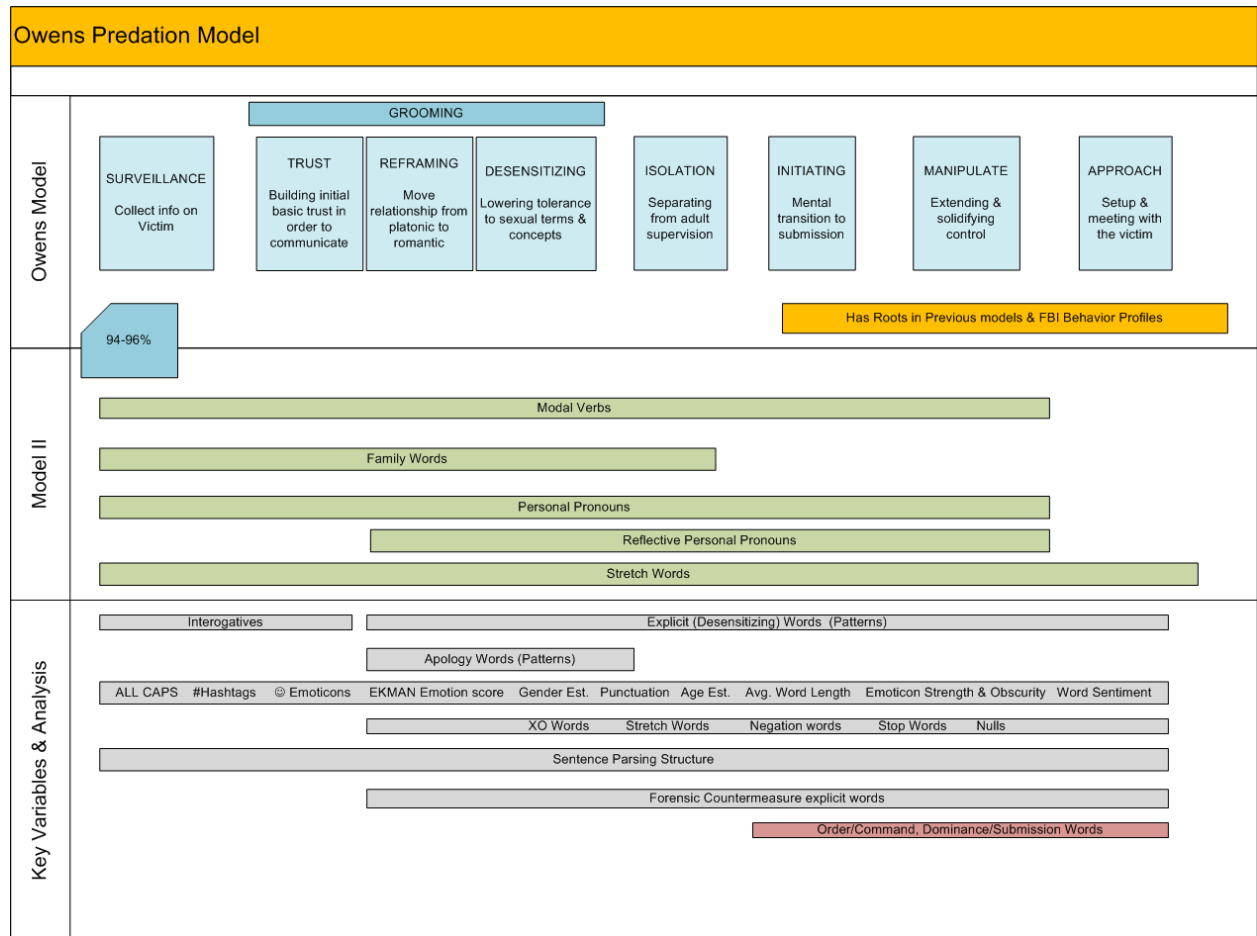


Figure 2: Proposed Predation Model (Owens).

Models Phases of a Sexual Predator's Attack Sequence[1, 2, 9]

1. Collection Information about the victim. [Surveillance]
2. Lowering the Victim's inhibitions. [Grooming]
 - a. Trust, or building initial basic trust in order to communicate.
 - b. Reframing the relationship from platonic to romantic
 - c. Desensitizing the victim to the use of sexual terms
3. Isolating the victim from adult supervision [Isolation]
4. Initiating the abuse [Initiate]
5. Extending control and influence [Manipulate]
6. Attempting to meet with the victim [Approach]

Metric indicators of Predators

1. Question words or Interrogatives, are word used to gather information.
2. ALLCAPS: the number of Messages with all characters in uppercase;
3. Hashtags: the number of hashtags;
4. Negation: the number of negated contexts. A negated context also affects the ngram and lexicon features: each word and associated with it polarity in a negated context become negated (e.g., 'not perfect' becomes 'not perfect_NEG', 'POLARITY_positive' becomes 'POLARITY_positive_NEG');
5. Relationship words like “meet”, “hookup”, “boyfriend” and “girlfriend” typically are overused by sexual predators. [6]
6. Family Words are used at an increased rate, as the predator attempts to separate the victim from their family by pointing out differences between them and their family, in addition to seeking information about how strong the victim's relationship is to their family.
7. Personal Pronouns get used by sexual predators at an increased rate, very similar to relation and family words. McGhee [2] indicated that splitting up this variable into 1st, 2nd and 3rd person pronouns and combining them with other word variables are also indicators or predatory behaviour.
8. Overuse of Reflexive Pronouns like “myself” or “yourself”. Again these get overused during the grooming process to draw distinction between the victim and others.
9. XOXO count. The use of “xo” or “XO” to mean hugs and kisses is also a verbal grooming by sexual predators. This introduces the idea of being touched and interacting sexually with the predator is gets used during the grooming phase of the attack.
10. Modal Verbs, is a type of auxiliary verb that is used to indicate a likelihood, ability, permission, and obligation. Examples include the English verbs can, could, may, might, must, will/, would, and, shall/, should.
11. Stretch words are words that are elongated for emphasis, typically used by the predator to sound more like a child in their use of words and lingo as a way to bond with the victim. For example, Noooooooooooo!
12. #HASHTAB usage by predators also tends to be high. Similar to Stretch words, this variable also may not be an indicator by itself and should be used with other variables or with an Age indicator variable.
13. Negation of words is used to sound more positive and is a common word trick used by all types of online predators, not just sexual predators. This variable is also typically used in combination with other variables.
14. Affect Word Score.
15. Desensitizing words usage.
16. Apology Word usage

17. Sentiment word score.
18. Emoticons:
 - a. presence/absence of positive and negative emoticons at any position in the message
 - b. whether the last token is a positive or negative emoticon.
 - c. Use of emoticon only as the entire message post.
 - d. Emotional sentiment of the emoticon.
 - e. Chiral or handedness of the emoticon
19. Punctuation:
 - a. The number of contiguous sequences of exclamation marks, question marks, and both exclamation and question marks; whether a term contains punctuation sequences such as "?!" and "!!!"
 - b. Whether the last token contains exclamation or question mark;
20. Stop Words
 - a. The Count in a message.
 - b. The percentage of message posts that contain only stop words.
21. Average word Length is shorter for sexual predators than normal adults as they try to use text lingo to "talk like children" and relate better to their victims.
22. Forensic Countermeasure Desensitizing words. This is a theory of the authors that sexual predators use desensitizing words by mis-spelling other words to combine the use. This functions as a forensic countermeasure. Examples are words like "Welcum" that incorporate an explicit word into another word to both disguise and desensitize. A normalized count if these words is used.

Metric Indicator Discussion & Code Logic

Data Cleanup and Conditioning

For the purpose of this paper the data feed is assumed to be Blog posts, IM Feeds, Twitter feeds, AOL Feeds, or other similar feeds. Nearly all of these types of systems have some common data elements, but the process of bringing multiple feeds together and merging them will take some customized code to achieve. We will discuss some basic example coding for getting the data into a SAS data format. We will discuss starting with either an XML file format feed or a single field CSV format, but the key data contained in the field will include

1. Date:Time (Can be a single DateTime or separate Date and Time)
2. Username
3. Message

A typical line of incoming XML format would look like this:

<date>2014-02-14</date><time>08:16:19</time><user>minkks</user><msg>OH MY GOODNESS</msg>

If your organization has licensed the SAS XML MAPPER tool, that is the easiest way to bring in XML data. The tool allows you to import the defined fields and define an output SAS dataset. Properties in each of the defined fields can be retained and brought into the SAS server.

If you do not have the SAS XML MAPPER tool, you can bring in the XML file as a single field text or CSV feed / file and parse it into the defined fields that you need. The PROC SQL function works well for this task.

Single field CSV feeds from Instant Message (IM) systems typically look like this.

<Name>minkks<date-time>2014-02-14:08:16:19<Message>OH MY GOODNESS

An example of the Import and parsing code used by the author is below.

/* Code for importing the single field data using a DATA step from "ChatLog" */

```
DATA WORK.ChatLog;
  LENGTH
    F1          $ 513 ;
  FORMAT
    F1          $CHAR513. ;
  INFORMAT
    F1          $CHAR513. ;
  INFILE '<path name for the data file or data feed>'
    LRECL=32767
    ENCODING="LATIN1"
    DLM='09'x
    MISSEVER
    DSD ;
  INPUT
    F1          : $CHAR513. ;
RUN;
```

/* Code for Parsing a single field feed into NAME, DATETIME and MESSAGE fields */

```
PROC SQL;
  CREATE TABLE WORK.QUERY_FOR_CHATLOG AS
  SELECT t1.F1,
    /* NAME */
    (STRIP(SUBSTR(t1.F1,1,(INDEX(t1.F1,""))))) LABEL="NAME" AS NAME,
    /* DATETIME */
```

```

        (SUBSTR(t1.F1, INDEX(t1.F1,"")+1, 20)) LABEL="DATETIME" AS DATETIME,
/* MESSAGE */
        (STRIP(SUBSTR(t1.F1,(INDEX(t1.F1,"")+2)))) LABEL="MESSAGE" AS MESSAGE
FROM WORK.CHATLOG t1;
QUIT;

```

If the user has a need, additional PROC SQL code can be added to further parse the DATETIME field into a separate DATE and TIME field.

```

/* PROC SQL code for separating DATETIME into DATE and TIME */
PROC SQL;
    CREATE TABLE WORK.QUERY_FOR_CHATLOG_0000 AS
    SELECT t1.F1,
           t1.NAME,
           t1.DATETIME,
/* DATE */
           (INPUT(SUBSTR(t1.DATETIME, 1, 8), MMDDYY10.)) FORMAT=MMDDYY10.
    LABEL="DATE" AS DATE,
/* TIME */
           (INPUT(SUBSTR(t1.DATETIME, 10, 8), IS8601TM8.)) FORMAT=IS8601TM8.
    LABEL="TIME" AS TIME,
           t1.MESSAGE
    FROM WORK.QUERY_FOR_CHATLOG t1;
QUIT;

```

In the case of the authors system, the incoming data field was parsed into separate DATE and TIME fields. The balance of the paper will work code examples based on this, but the users should be able to substitute DATETIME fairly easily.

Establishing Baselines

Many of the metrics used to detect online predators function by comparing against a non-predator or a victim using normalized data. Therefore we need to have appropriate normalization values for our data. In some cases we will be focusing on how a predator uses individual graphs (letters or symbols) or characters, so a count of graphs in each post as well as in total is needed. We will also look at how they use words, so again, word count by individual post and in total are needed. The last normalization would be the total number of message posts or message posts per unit time.

All of these baselines can be created within a PROC SQL command against our message log data. The following code does this.

```

PROC SQL;

```

```

CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
SELECT t1.F1, /* F1 is the Message index number. */
       t1.NAME,
       t1.DATE,
       t1.TIME,
       t1.MESSAGE,
       /* WORDCOUNT */
       (COUNTW(t1.MESSAGE)) LABEL="WORDCOUNT" AS WORDCOUNT,
       /* LETTERCOUNT */
       (LENGTHN(t1.MESSAGE)) LABEL="LETTERCOUNT" AS LETTERCOUNT,
       /* MESSAGE COUNT */
       (COUNT(1) LABEL="MESSAGECOUNT" AS MESSAGECOUNT,
QUIT;

```

The above code in combination with summation by the t1.NAME variable will be used for normalizing the various indexes used in the program. Most of the following index calculations can be added to the same PROC SQL statement as above right after the MESSAGECOUNT code.

QUESTION WORDS (INTERROGATIVES)

Interrogatives for the basis for questions within the English language and are used by people to request and gather information from others during a conversation. For Sexual predators, this is vital for Step 1 of the Attack Sequence and failure here stops the entire process. Inclusive in this list is the standard ones, who, what, when, where, why, whom, which and how. With the advent of internet slang, a larger list of words, abbreviation and symbols have entered the vernacular that function as interrogatives. This include ?, dig, 3rd degree, 3rd, third degree, fish, grill, hammer, pimp, pump, d, D. The d and D are short for "Details" and get used as follows

VICTIM: I am going out with a friend tonight.
 PREDATOR: d

VICTIM: I am going out with a friend tonight.
 PREDATOR: D

Words like Fish and Hammer can also have Emoticon substitutes which may also be added to the list at the programmer's discretion. The variable becomes a count of Interrogatives and can be normalized in the same manner as other variables. A corpus of Interrogatives is listed in Appendix E and the reader is also referred to section on Forensic Countermeasures.

ALL CAPS

The use of ALL CAPS in the entire word post is the real world equivalent of shouting. In terms of a predator, this is an emotional outburst, which sexual predators are known to use often due to emotional immaturity. A simple Yes/No Count of the number of posts using ALL CAPS is one indicator of a sexual predator and can be used in conjunction with other variables for detection. In terms of SAS code we can use the following to create an ALLCAPS indicator variable. We make use of the ANYLOWER function to detect if the MESSAGE string has any lower case letters and return the first position of the lowercase letter. Any returns would be greater than 1, so we use the IFN function to flip the 1, 0 results around and output an ALLCAPS = 1 if the MESSAGE is ALL CAPS and a 0 if it is not.

```
/* ALLCAPS Function to determine if the message is all caps or not. */  
(IFN(ANYLOWER(t1.MESSAGE)=0,1,0)) LABEL="ALLCAPS" AS ALLCAPS  
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
```

Hashtags

The total number of hashtags used is also an indicator of child sexual predators. The predator tends to over-use hashtags in an attempt to talk like a child. We'll typically see normal adults using hashtags at some level per post or per 1000 words, we will see children using Hashtags at a higher level, and sexual predators at a higher level still. The actual numerical levels will vary depending on the forum or site, so this detection is comparing the level of use between the predator and the victim or the predator and other adults of similar age.

To create a variable from our MESSAGE posts called HASHTAG, we can use SAS code like this.

```
/* HASHTAG Function to detect the use of a Hashtag in a MESSAGE post. */  
(IFN(t1.MESSAGE CONTAINS "#", 1,0) ) LABEL="HASTAG" AS HASHTAG  
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
```

Negation

The increased use of the word "Not" for negation to modify other words is a characteristic of sexual predators. By using positive words and negating them, the predator attempts to stay "more approachable" and "open" and avoid appearing "pessimistic" or "negative" to their victims. Normalized comparisons can be made against victims or other adults to identify predators. To create a NEGATION variable from a MESSAGE post, the following SAS code can be used.

```
/* NEGATION */  
(IFN(INDEXW(t1.MESSAGE, "not")>=1, 1, 0) ) LABEL="NEGATION" AS NEGATION
```

```
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
```

Relation

The overuse of relation words during the grooming phase of a sexual predators online attack can also be used to identify the predator. Relation words like “meet”, “hookup”, “boyfriend” and “girlfriend” typically are overused by sexual predators. [6] A corpus of relation words for english have been developed and used create the RELATIONWORD variable from the MESSAGE field. This is a count of the number of such words in each message post. This variable can be aggregated in total as well as compared by trend usage during different attack phases as indicators of a sexual predator. The SAS code for creating this variable is below.

```
PROC SQL;
```

```
CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
SELECT t1.F1,
       t1.NAME,
       t1.DATE,
       t1.TIME,
       t1.MESSAGE,
       /* RelationWord */
       (COUNT(t1.MESSAGE, "meet", 'i') + COUNT(t1.MESSAGE, "date", 'i') +
COUNT(t1.MESSAGE, "boyfriend", 'i') +
COUNT(t1.MESSAGE, "girlfriend", 'i') + COUNT(t1.MESSAGE, "hookup", 'i') +
COUNT(t1.MESSAGE, "hang", 'i')
       ) + COUNT(t1.MESSAGE, "acquainted", 'i') + COUNT(t1.MESSAGE, "affiliated", 'i')
+ COUNT(t1.MESSAGE, "arm's-length", 'i') + COUNT(t1.MESSAGE, "brittle", 'i')
+ COUNT(t1.MESSAGE, "broken", 'i') + COUNT(t1.MESSAGE, "bromantic", 'i')
+ COUNT(t1.MESSAGE, "brotherly", 'i') + COUNT(t1.MESSAGE, "chummy", 'i')
+ COUNT(t1.MESSAGE, "clannish", 'i') + COUNT(t1.MESSAGE, "close", 'i')
+ COUNT(t1.MESSAGE, "close", 'i') + COUNT(t1.MESSAGE, "close", 'i')
+ COUNT(t1.MESSAGE, "close", 'i') + COUNT(t1.MESSAGE, "connected", 'i')
+ COUNT(t1.MESSAGE, "cosy", 'i') + COUNT(t1.MESSAGE, "cozy", 'i')
+ COUNT(t1.MESSAGE, "dysfunctional", 'i') + COUNT(t1.MESSAGE, "estranged", 'i')
+ COUNT(t1.MESSAGE, "fragile", 'i') + COUNT(t1.MESSAGE, "fraternal", 'i')
+ COUNT(t1.MESSAGE, "fraternal", 'i') + COUNT(t1.MESSAGE, "friendly", 'i')
+ COUNT(t1.MESSAGE, "go", 'i') + COUNT(t1.MESSAGE, "have", 'i')
+ COUNT(t1.MESSAGE, "have", 'i') + COUNT(t1.MESSAGE, "heavy", 'i')
+ COUNT(t1.MESSAGE, "illicit", 'i') + COUNT(t1.MESSAGE, "immediate", 'i')
+ COUNT(t1.MESSAGE, "inseparable", 'i') + COUNT(t1.MESSAGE, "interpersonal", 'i')
+ COUNT(t1.MESSAGE, "intimate", 'i') + COUNT(t1.MESSAGE, "intimate", 'i')
+ COUNT(t1.MESSAGE, "intimate", 'i') + COUNT(t1.MESSAGE, "intimately", 'i')
+ COUNT(t1.MESSAGE, "long-lost", 'i') + COUNT(t1.MESSAGE, "loveless", 'i')
+ COUNT(t1.MESSAGE, "maternal", 'i') + COUNT(t1.MESSAGE, "matrilineal", 'i')
+ COUNT(t1.MESSAGE, "monogamous", 'i') + COUNT(t1.MESSAGE, "monogamously", 'i')
```

```

+ COUNT(t1.MESSAGE, "mouth" , 'i' )+ COUNT(t1.MESSAGE, "one-sided" , 'i' )
+ COUNT(t1.MESSAGE, "one-to-one" , 'i' )+ COUNT(t1.MESSAGE, "one-way" , 'i' )
+ COUNT(t1.MESSAGE, "patriarchal" , 'i' )+ COUNT(t1.MESSAGE, "patrilineal" , 'i' )
+ COUNT(t1.MESSAGE, "personal" , 'i' )+ COUNT(t1.MESSAGE, "personally" , 'i' )
+ COUNT(t1.MESSAGE, "platonic" , 'i' )+ COUNT(t1.MESSAGE, "platonically" , 'i' )
+ COUNT(t1.MESSAGE, "political" , 'i' )+ COUNT(t1.MESSAGE, "polyandrous" , 'i' )
+ COUNT(t1.MESSAGE, "polygamous" , 'i' )+ COUNT(t1.MESSAGE, "related" , 'i' )
+ COUNT(t1.MESSAGE, "rocky" , 'i' )+ COUNT(t1.MESSAGE, "same-sex" , 'i' )
+ COUNT(t1.MESSAGE, "serious" , 'i' )+ COUNT(t1.MESSAGE, "sexual" , 'i' )
+ COUNT(t1.MESSAGE, "shifting" , 'i' )+ COUNT(t1.MESSAGE, "strong" , 'i' )
+ COUNT(t1.MESSAGE, "suited" , 'i' )+ COUNT(t1.MESSAGE, "symbiotic" , 'i' )
+ COUNT(t1.MESSAGE, "thick" , 'i' )+ COUNT(t1.MESSAGE, "tight" , 'i' )
+ COUNT(t1.MESSAGE, "tightknit" , 'i' )+ COUNT(t1.MESSAGE, "unstable" , 'i' )
+ COUNT(t1.MESSAGE, "warming" , 'i' )+ COUNT(t1.MESSAGE, "a hungry mouth" , 'i' )
+ COUNT(t1.MESSAGE, "a hungry mouth to feed" , 'i' )+ COUNT(t1.MESSAGE, "an old friend" ,
'i' )+ COUNT(t1.MESSAGE, "an old ally" , 'i' )+ COUNT(t1.MESSAGE, "an old enemy" , 'i' )
+ COUNT(t1.MESSAGE, "an old student" , 'i' )+ COUNT(t1.MESSAGE, "an old girlfriend" , 'i' )
+ COUNT(t1.MESSAGE, "thick as thieves" , 'i' )+ COUNT(t1.MESSAGE, "at arm's length" , 'i' )
+ COUNT(t1.MESSAGE, "at arms length" , 'i' )+ COUNT(t1.MESSAGE, "be on good terms" , 'i' )
+ COUNT(t1.MESSAGE, "be on bad terms" , 'i' )+ COUNT(t1.MESSAGE, "be on friendly terms"
, 'i' )+ COUNT(t1.MESSAGE, "get along famously" , 'i' )+ COUNT(t1.MESSAGE, "get on
famously" , 'i' )+ COUNT(t1.MESSAGE, "not on speaking terms" , 'i' )+ COUNT(t1.MESSAGE,
"on the good side of" , 'i' )+ COUNT(t1.MESSAGE, "on the bad side of" , 'i' )
+ COUNT(t1.MESSAGE, "on the right side of" , 'i' )+ COUNT(t1.MESSAGE, "on the wrong side
of" , 'i' )+ COUNT(t1.MESSAGE, "nodding acquaintance" , 'i' )+ COUNT(t1.MESSAGE, "nodding
terms" , 'i' )+ COUNT(t1.MESSAGE, "the best of friends" , 'i' )) LABEL="RelationWord" AS
RelationWord,
    FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;

```

Family Words

In a pattern similar to RELATION words, FAMILY words are also used at an increased rate by sexual predators. Predators use family references to gather information about the victim's relationship to family members in order to learn information about the victim, like how close a victim is to their parents and if they are likely to confide in them; These words are also used by the predator show the victim how different they are from their family in order to induce the victim to emotionally separate from their family and the inherent protection they offer.

Again, FAMILYWORD index is a total count of the number of usages of these words and it will need to be normalized. The SAS code to generate this variable is below.

```
PROC SQL;
```



```

CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
SELECT t1.F1,
       t1.NAME,
       t1.DATE,
       t1.TIME,
       t1.MESSAGE,
       /* FamilyWords */
       (+ COUNT(t1.MESSAGE, "mom", 'i' )
+ COUNT(t1.MESSAGE, "father", 'i' )
+ COUNT(t1.MESSAGE, "dad", 'i' )
+ COUNT(t1.MESSAGE, "parent", 'i' )
+ COUNT(t1.MESSAGE, "children", 'i' )
+ COUNT(t1.MESSAGE, "son", 'i' )
+ COUNT(t1.MESSAGE, "daughter", 'i' )
+ COUNT(t1.MESSAGE, "sister", 'i' )
+ COUNT(t1.MESSAGE, "brother", 'i' )
+ COUNT(t1.MESSAGE, "grandmother", 'i' )
+ COUNT(t1.MESSAGE, "grandfather", 'i' )
+ COUNT(t1.MESSAGE, "grandparent", 'i' )
+ COUNT(t1.MESSAGE, "grandson", 'i' )
+ COUNT(t1.MESSAGE, "granddaughter", 'i' )
+ COUNT(t1.MESSAGE, "grandchild", 'i' )
+ COUNT(t1.MESSAGE, "aunt", 'i' )
+ COUNT(t1.MESSAGE, "uncle", 'i' )
+ COUNT(t1.MESSAGE, "niece", 'i' )
+ COUNT(t1.MESSAGE, "nephew", 'i' )
+ COUNT(t1.MESSAGE, "cousin", 'i' )
+ COUNT(t1.MESSAGE, "husband", 'i' )
+ COUNT(t1.MESSAGE, "wife", 'i' )
+ COUNT(t1.MESSAGE, "sister-in-law", 'i' )
+ COUNT(t1.MESSAGE, "brother-in-law", 'i' )
+ COUNT(t1.MESSAGE, "mother-in-law", 'i' )
+ COUNT(t1.MESSAGE, "father-in-law", 'i' )
+ COUNT(t1.MESSAGE, "partner", 'i' )
+ COUNT(t1.MESSAGE, "fiancé", 'i' )
+ COUNT(t1.MESSAGE, "fiancée", 'i' )
+ COUNT(t1.MESSAGE, "fiance", 'i' )
+ COUNT(t1.MESSAGE, "fiancee", 'i' )
+ COUNT(t1.MESSAGE, "sis", 'i' )
+ COUNT(t1.MESSAGE, "mum", 'i' )
+ COUNT(t1.MESSAGE, "cuz", 'i' )
+ COUNT(t1.MESSAGE, "bro", 'i' )
+ COUNT(t1.MESSAGE, "pop", 'i' ) ) LABEL="FamilyWords" AS FamilyWords,

```

```
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;
```

Personal Pronouns

Overuse of personal pronouns as a bonding method is common with sexual predators. A comparison needs to be drawn between victims and normal forum users and the predators on any particular venue as the level of absolute usage can vary. It is the delta between the normal user and the predator that we look for with this variable. The SAS code for the creation of PERSPRONOUN from MESSAGE.

```
PROC SQL;
CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
SELECT t1.F1,
       t1.NAME,
       t1.DATE,
       t1.TIME,
       t1.MESSAGE,
       /* PersPronoun */
       (COUNT(t1.MESSAGE, "I ") + COUNT(t1.MESSAGE, "me ", 'i') +
COUNT(t1.MESSAGE, "my ", 'i') +
COUNT(t1.MESSAGE, "mine ", 'i') + COUNT(t1.MESSAGE, "you", 'i') +
COUNT(t1.MESSAGE, "your ", 'i') +
COUNT(t1.MESSAGE, "yours", 'i') + COUNT(t1.MESSAGE, "he ", 'i') +
COUNT(t1.MESSAGE, "she ", 'i') +
COUNT(t1.MESSAGE, " it ", 'i') + COUNT(t1.MESSAGE, "him", 'i') +
COUNT(t1.MESSAGE, "his", 'i') +
COUNT(t1.MESSAGE, "her", 'i') + COUNT(t1.MESSAGE, "its", 'i') +
COUNT(t1.MESSAGE, "ours", 'i') +
COUNT(t1.MESSAGE, "they", 'i') + COUNT(t1.MESSAGE, "hers", 'i') +
COUNT(t1.MESSAGE, "we ", 'i') +
COUNT(t1.MESSAGE, " us ", 'i') + COUNT(t1.MESSAGE, "our ", 'i') +
COUNT(t1.MESSAGE, "them", 'i') +
COUNT(t1.MESSAGE, "their", 'i') + COUNT(t1.MESSAGE, "theirs", 'i') +
COUNT(t1.MESSAGE, "u ", 'i') +
COUNT(t1.MESSAGE, "u?", 'i')) LABEL="PersPronoun" AS PersPronoun,
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;
```

A list of English Personal pronouns is here. [I me my mine you your yours
he she it him his her its hers we us our ours they them their
theirs and the slang version “u” and “u?”]

McGhee [2] indicated in his results that separating the pronouns into 1st, 2nd and 3rd person and tracking each as a separate variable increased the power of detection of sexual predators. He offered no details on how they were actually used in the program code, but I suspect it was in combination with other variables within a given post. There may also be value in counting pronouns in the Case Dimension (SUBJECTIVE, OBJECTIVE, POSSESSIVE) and in the PLURALITY Dimension (SINGULAR, PLURAL).

A Personal_Pronoun_Corpus is included in APPENDIX

Objective Reflexive Pronouns

Overuse of Reflexive Pronouns like “myself” or “yourself”. Again these get overused during the grooming process to draw distinction between the victim and others.

```
PROC SQL;
  CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
  SELECT t1.F1,
         t1.NAME,
         t1.DATE,
         t1.TIME,
         t1.MESSAGE,
         /* REFPrnouns, also contain archaic Logophors */
         (COUNT(t1.MESSAGE, "myself", 'i') + COUNT(t1.MESSAGE, "yourself", 'i')
+ COUNT(t1.MESSAGE, "thyself", 'i')+ COUNT(t1.MESSAGE, "himself", 'i')
+ COUNT(t1.MESSAGE, "hissel", 'i')+ COUNT(t1.MESSAGE, "herself", 'i')
+ COUNT(t1.MESSAGE, "itself", 'i')+ COUNT(t1.MESSAGE, "oneself", 'i')
+ COUNT(t1.MESSAGE, "ourselves", 'i')+ COUNT(t1.MESSAGE, "ourself", 'i')
+ COUNT(t1.MESSAGE, "yourselves", 'i')+ COUNT(t1.MESSAGE, "themselves", 'i')
+ COUNT(t1.MESSAGE, "theirselves", 'i')+ COUNT(t1.MESSAGE, "theirselves", 'i')
) LABEL="REFPrnouns" AS REFPrnouns,
  FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;
```

XO Count

The use of “xo” or “XO” to mean hugs and kisses is also a form of verbal grooming by sexual predators. This introduces the idea of being touched and interacting sexually with the predator and is used during the grooming phase of the attack. Extended variant so this like “xoxoxoxoxo” are common and can be detected by a simple search for “xo”, which occurs infrequently enough in regular english usage that the error induced typically falls to zero. Like the measures above, a normalized comparison can be used, but predators often stand out in a pure count as well.

SAS code for use inside of a PROC SQL command follows.

```
PROC SQL;
  CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
  SELECT t1.F1,
         t1.NAME,
         t1.DATE,
         t1.TIME,
         t1.MESSAGE,
         /* XOCOUNT */
         (COUNT(t1.MESSAGE, "xo", 'i')) LABEL="XOCOUNT" AS XOCOUNT,
  FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;
```

Modal Verbs

Modal Verbs, is a type of auxiliary verb that is used to indicate a likelihood, ability, permission, and obligation. Examples include the English verbs can, could, may, might, must, will/, would, and, shall/, should. These types of words are used by predators to obligate the victim to certain actions, trick them into giving permission, control their actions, or transfer responsibility of some action back to the victim. These words get overused by manipulative predators and stand out easily. This variable is usually a significant variable in any detection model of predators in general and sexual predators especially.

SAS code to calculate MODALVERB from MESSAGE inside a PROC SQL statement is below.

```
PROC SQL;
  CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
  SELECT t1.F1,
         t1.NAME,
         t1.DATE,
         t1.TIME,
         t1.MESSAGE,
         /* ModalVerbs */
         (COUNT(t1.MESSAGE, "can", 'i') + COUNT(t1.MESSAGE, "could", 'i') +
COUNT(t1.MESSAGE, "may", 'i') +
COUNT(t1.MESSAGE, "might", 'i') + COUNT(t1.MESSAGE, "shall", 'i') +
COUNT(t1.MESSAGE, "shall", 'i') +
COUNT(t1.MESSAGE, "should", 'i') + COUNT(t1.MESSAGE, "will", 'i') +
COUNT(t1.MESSAGE, "would", 'i') +
COUNT(t1.MESSAGE, "must", 'i') + COUNT(t1.MESSAGE, "ought", 'i') +
COUNT(t1.MESSAGE, "dare", 'i') +
```

```

COUNT(t1.MESSAGE, "need", 'i') + COUNT(t1.MESSAGE, "darest", 'i') +
COUNT(t1.MESSAGE, "had better", 'i') +
COUNT(t1.MESSAGE, "used to", 'i')) LABEL="ModalVerbs" AS ModalVerbs,
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;

```

Stretch Index or Stretch Words

Stretch words are words that are elongated for emphasis, typically used by the predator to sound more like a child in their use of words and lingo as a way to bond with the victim. For example, Nooooooooooooo! Often this variable would need to be combined with some type of Age index or measurement in order to be an indicator of a predator. Children and younger people often use this type of word play for emphasis on short message systems, so unless it is combined with an age indicator, or other variables, it is not an indicator on its own.

The SAS code for calculating the STRETCH_INDEX from MESSAGE is below.

```

PROC SQL;
CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
SELECT t1.F1,
       t1.NAME,
       t1.DATE,
       t1.TIME,
       t1.MESSAGE,
       /* Stretch_Index */
       (COUNT(t1.MESSAGE, "aaa", "i")+ COUNT(t1.MESSAGE, "bbb", "i")+
COUNT(t1.MESSAGE, "ccc", "i")+
COUNT(t1.MESSAGE, "ddd", "i")+ COUNT(t1.MESSAGE, "eee", "i")+
COUNT(t1.MESSAGE, "fff", "i")+
COUNT(t1.MESSAGE, "ggg", "i")+ COUNT(t1.MESSAGE, "hhh", "i")+
COUNT(t1.MESSAGE, "iii", "i")+
COUNT(t1.MESSAGE, "jjj", "i")+ COUNT(t1.MESSAGE, "kkk", "i")+
COUNT(t1.MESSAGE, "lll", "i")+
COUNT(t1.MESSAGE, "mmm", "i")+ COUNT(t1.MESSAGE, "nnn", "i")+
COUNT(t1.MESSAGE, "ooo", "i")+
COUNT(t1.MESSAGE, "ppp", "i")+ COUNT(t1.MESSAGE, "qqq", "i")+
COUNT(t1.MESSAGE, "rrr", "i")+
COUNT(t1.MESSAGE, "sss", "i")+ COUNT(t1.MESSAGE, "ttt", "i")+
COUNT(t1.MESSAGE, "uuu", "i")+
COUNT(t1.MESSAGE, "vvv", "i")+ COUNT(t1.MESSAGE, "www", "i")+
COUNT(t1.MESSAGE, "xxx", "i")+
COUNT(t1.MESSAGE, "yyy", "i")+ COUNT(t1.MESSAGE, "zzz", "i")+
COUNT(t1.MESSAGE, "111", "i")+

```

```

COUNT(t1.MESSAGE, "222", "i")+ COUNT(t1.MESSAGE, "333", "i")+
COUNT(t1.MESSAGE, "444", "i")+
COUNT(t1.MESSAGE, "555", "i")+ COUNT(t1.MESSAGE, "666", "i")+
COUNT(t1.MESSAGE, "777", "i")+
COUNT(t1.MESSAGE, "888", "i")+ COUNT(t1.MESSAGE, "999", "i")+
COUNT(t1.MESSAGE, "000", "i")+
COUNT(t1.MESSAGE, "``", "i")+ COUNT(t1.MESSAGE, "~~", "i")+
COUNT(t1.MESSAGE, "!!!", "i")+
COUNT(t1.MESSAGE, "@@@", "i")+ COUNT(t1.MESSAGE, "###", "i")+
COUNT(t1.MESSAGE, "$$$", "i")+
COUNT(t1.MESSAGE, "%%%", "i")+ COUNT(t1.MESSAGE, "^^^", "i")+
COUNT(t1.MESSAGE, "&&&", "i")+
COUNT(t1.MESSAGE, "***", "i")+ COUNT(t1.MESSAGE, "(((", "i")+
COUNT(t1.MESSAGE, ")))", "i")+
COUNT(t1.MESSAGE, "===", "i")+ COUNT(t1.MESSAGE, "+++", "i")+
COUNT(t1.MESSAGE, "\\\\", "i")+
COUNT(t1.MESSAGE, "|||", "i")+ COUNT(t1.MESSAGE, "[[[", "i")+
COUNT(t1.MESSAGE, "]]]", "i")+
COUNT(t1.MESSAGE, "}}}", "i")+ COUNT(t1.MESSAGE, "{{{", "i")+
COUNT(t1.MESSAGE, "", "i")+ COUNT(t1.MESSAGE,
";;;", "i")+ COUNT(t1.MESSAGE, ":::", "i")+ COUNT(t1.MESSAGE, "???", "i")+
COUNT(t1.MESSAGE, "///", "i")+
COUNT(t1.MESSAGE, "...", "i")+ COUNT(t1.MESSAGE, ",,,", "i")+
COUNT(t1.MESSAGE, "<<<", "i")+
COUNT(t1.MESSAGE, ">>>", "i")) LABEL="Stretch_Index" AS Stretch_Index,
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;

```

#Hashtags

#HASHTAG usage by predators also tends to be high. Similar to Stretch words, this variable also may not be an indicator by itself and should be used with other variables or with an Age indicator variable.

```

PROC SQL;
CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
SELECT t1.F1,
       t1.NAME,
       t1.DATE,
       t1.TIME,
       t1.MESSAGE,
       /* HASHTAG */
       (IFN(t1.MESSAGE CONTAINS "#", 1,0) ) LABEL="HASTAG" AS HASHTAG,

```

```
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;  
QUIT;
```

Negation Usage

Negation of words is used to sound more positive and is a common word trick used by all types of online predators, not just sexual predators. This variable is typically used in combination with other variables.

```
PROC SQL;  
CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS  
SELECT t1.F1,  
       t1.NAME,  
       t1.DATE,  
       t1.TIME,  
       t1.MESSAGE,  
       /* NEGATION */  
       (IFN(INDEXW(t1.MESSAGE, "not")>=1, 1, 0) ) LABEL="NEGATION" AS NEGATION  
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;  
QUIT;
```

AFFECT Word Score

The Affect word score of a message post, Message thread or a Message conversation is broken out along the dimensions of Pleasantness, Activation, Imagery, and total Affect. These dimensions can be obtained by scoring each word in a message post, summing the respective scores and using them as variables for comparing predators to other message posters. Because sexual predators exhibit emotional immaturity, both the average scores and the variation of the scores can be used to distinguish them from a normal user.

These calculations require the importation of the Dictionary of Affect and Joining it with the MESSAGE data via table join commands. For this SAS code we bring in the CHATLOG (with most of the previous index variables) and the cleaned version of the DICTIONARYOFAFFECTCLEAN and join them. We also bring in the Pleasantness, Activation, Imagery,scores from the dictionary. The code is below.

```
PROC SQL;  
CREATE TABLE WORK.QUERY_FOR_CHATLOG_0002 AS  
SELECT t1.F1,  
       t1.NAME,  
       t1.DATE,  
       t1.TIME,  
       t1.MESSAGE,
```

```

t1.WORDCOUNT,
t1.RelationWord,
t1.FamilyWords,
t1.PersPronoun,
t1.REFPronouns,
t1.XOCount,
t1.ModalVerbs,
t1.Stretch_Index,
t2.aa,
t2.ii,
/* AFFECT */
(IFN( t1.MESSAGE CONTAINS t2.Word, 1,0)) LABEL="AFFECT" AS AFFECT
FROM WORK.QUERY_FOR_CHATLOG_0001 t1
RIGHT JOIN WORK.DICTIONARYOFAFFECTCLEAN t2 ON (t1.MESSAGE = t2.Word);
QUIT;
/* Summarization tables for above AFFECT scores */
TITLE;
TITLE1 "Summary Tables";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL) on
%TRIM(%QSYSFUNC(DATE()), NLDATE20.) at %TRIM(%SYSFUNC(TIME()),
TIMEAMPM12.)";
PROC TABULATE
DATA=WORK.QUERY_FOR_CHATLOG_0002;

VAR ee aa ii AFFECT;
CLASS NAME / ORDER=UNFORMATTED MISSING;
TABLE /* Row Dimension */
NAME,
/* Column Dimension */
Sum*(
ee
aa
ii
AFFECT) ;
;
RUN;
RUN; QUIT;
TITLE; FOOTNOTE;

```


Desensitizing word usage

During the initial approach phase, a sexual predator will use non-explicit words so they don't alarm their victim. They want the child to first get use to talking to them but once that happens they want to start a process called "Desensitization". This is the process where the predator gets the victim conditioned to the idea of having sexual relations by getting them to accept the use of talking with sexually explicit words.

The predator will start by working in a few explicit words into the conversation. If the victim objects or "calls them out" on the use of those words, the predator will usually respond with 1. an apology or 2. will chastise them for being "a baby" or "a little kid", 3. challenge them to "grow up". If the apology option is use, the predator will continue to use explicit words and just apologize each time until the victim stops challenging them on the word usage. In any of these 3 scenarios, the predator will ramp up the use of desensitizing words as the conversation progresses.

If apology words [7] are used with the desensitizing words, they will ramp up initially with them, but at some point will fall off after the victim stops any challenges. These two word types can form a clear pattern that can also be detected as a signature of a sexual predator. This pattern is illustrated in the graph below.

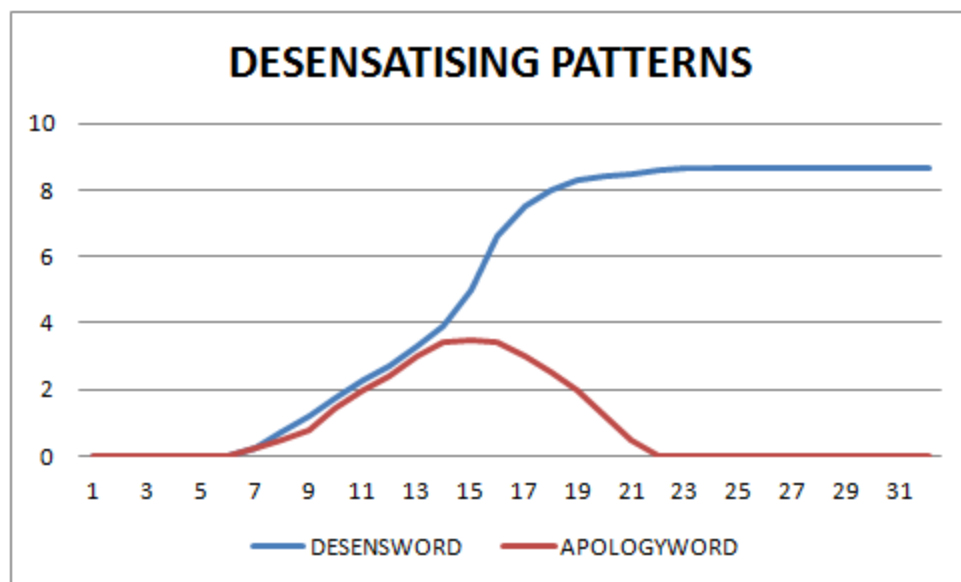


FIGURE 3: Desensitising and Apology Word use pattern.

There can be a lot of explicit words used for desensitizing victims, along with slang variations of those words. Desensitization may also take place across racial or gender boundaries and explicit racial or gender slurs or compliments may be used as well, all depending on the race and preferences of the sexual predator. It may be necessary to add or delete racial or gender

explicit words from the corpus depending on the social media site subject in order to deal with these types of word properly.

I think it is important to note that in an early regression model that had an accuracy of 91%, desensitizing words did not become a statistically significant variable and was dropped from the model. The use of other word patterns were likely influenced by the presence of explicit desensitizing words, but DESENSWORD itself was not used in the model to detect predators.

SAS Code for DESENSWORD index, as a count of words in a post is Appendix C1. The Explicit Word Corpus in Appendix B is used.

Apology word usage

Apology words may be used along with explicit words as part of a desensitisation process by sexual predators, but are seldom used by psychological predators (aka Trolls) except in a mocking manner. Apology word fall into a number of categories that include Apology Words, Apology Antonyms, Acknowledgments, amends, defense, excuse, justification, and parody words. All of these words get used as various forms of apology. Counts or percentage use of these words for the basis for measurement variables. A full corpus of these words is in APPENDIX C.[7] and SAS code is in Appendix C2.

Stop Words

```
PROC SQL;
  CREATE TABLE WORK.QUERY_FOR_CHATLOG_0001 AS
  SELECT t1.F1,
         t1.NAME,
         t1.DATE,
         t1.TIME,
         t1.MESSAGE,
         /* STOPWORDS */
         (COUNT(t1.MESSAGE, "a", "i")+ COUNT(t1.MESSAGE, "able", "i")+
COUNT(t1.MESSAGE, "about", "i")+
         COUNT(t1.MESSAGE, "across", "i")+ COUNT(t1.MESSAGE, "after", "i")+
COUNT(t1.MESSAGE, "all", "i")+
         COUNT(t1.MESSAGE, "almost", "i")+ COUNT(t1.MESSAGE, "also", "i")+
COUNT(t1.MESSAGE, "am", "i")+
         COUNT(t1.MESSAGE, "among", "i")+ COUNT(t1.MESSAGE, "an", "i")+
COUNT(t1.MESSAGE, "and", "i")+
         COUNT(t1.MESSAGE, "any", "i")+ COUNT(t1.MESSAGE, "are", "i")+
COUNT(t1.MESSAGE, "as", "i")+
```

COUNT(t1.MESSAGE, "at", "i")+ COUNT(t1.MESSAGE, "be", "i")+
COUNT(t1.MESSAGE, "because", "i")+
COUNT(t1.MESSAGE, "been", "i")+ COUNT(t1.MESSAGE, "but", "i")+
COUNT(t1.MESSAGE, "by", "i")+
COUNT(t1.MESSAGE, "can", "i")+ COUNT(t1.MESSAGE, "cannot", "i")+
COUNT(t1.MESSAGE, "could", "i")+
COUNT(t1.MESSAGE, "dear", "i")+ COUNT(t1.MESSAGE, "did", "i")+
COUNT(t1.MESSAGE, "do", "i")+
COUNT(t1.MESSAGE, "does", "i")+ COUNT(t1.MESSAGE, "either", "i")+
COUNT(t1.MESSAGE, "else", "i")+
COUNT(t1.MESSAGE, "ever", "i")+ COUNT(t1.MESSAGE, "every", "i")+
COUNT(t1.MESSAGE, "for", "i")+
COUNT(t1.MESSAGE, "from", "i")+ COUNT(t1.MESSAGE, "get", "i")+
COUNT(t1.MESSAGE, "got", "i")+
COUNT(t1.MESSAGE, "had", "i")+ COUNT(t1.MESSAGE, "has", "i")+
COUNT(t1.MESSAGE, "have", "i")+
COUNT(t1.MESSAGE, "he", "i")+ COUNT(t1.MESSAGE, "her", "i")+
COUNT(t1.MESSAGE, "hers", "i")+
COUNT(t1.MESSAGE, "him", "i")+ COUNT(t1.MESSAGE, "his", "i")+
COUNT(t1.MESSAGE, "how", "i")+
COUNT(t1.MESSAGE, "however", "i")+ COUNT(t1.MESSAGE, "i", "i")+
COUNT(t1.MESSAGE, "if", "i")+
COUNT(t1.MESSAGE, "in", "i")+ COUNT(t1.MESSAGE, "into", "i")+
COUNT(t1.MESSAGE, "is", "i")+
COUNT(t1.MESSAGE, "it", "i")+ COUNT(t1.MESSAGE, "its", "i")+
COUNT(t1.MESSAGE, "just", "i")+
COUNT(t1.MESSAGE, "least", "i")+ COUNT(t1.MESSAGE, "let", "i")+
COUNT(t1.MESSAGE, "like", "i")+
COUNT(t1.MESSAGE, "likely", "i")+ COUNT(t1.MESSAGE, "may", "i")+
COUNT(t1.MESSAGE, "me", "i")+
COUNT(t1.MESSAGE, "might", "i")+ COUNT(t1.MESSAGE, "most", "i")+
COUNT(t1.MESSAGE, "must", "i")+
COUNT(t1.MESSAGE, "my", "i")+ COUNT(t1.MESSAGE, "neither", "i")+
COUNT(t1.MESSAGE, "no", "i")+
COUNT(t1.MESSAGE, "nor", "i")+ COUNT(t1.MESSAGE, "not", "i")+
COUNT(t1.MESSAGE, "of", "i")+
COUNT(t1.MESSAGE, "off", "i")+ COUNT(t1.MESSAGE, "often", "i")+
COUNT(t1.MESSAGE, "on", "i")+
COUNT(t1.MESSAGE, "only", "i")+ COUNT(t1.MESSAGE, "or", "i")+
COUNT(t1.MESSAGE, "other", "i")+
COUNT(t1.MESSAGE, "our", "i")+ COUNT(t1.MESSAGE, "own", "i")+
COUNT(t1.MESSAGE, "rather", "i")+

```

COUNT(t1.MESSAGE, "said", "i")+ COUNT(t1.MESSAGE, "say", "i")+
COUNT(t1.MESSAGE, "says", "i")+
COUNT(t1.MESSAGE, "she", "i")+ COUNT(t1.MESSAGE, "should", "i")+
COUNT(t1.MESSAGE, "since", "i")+
COUNT(t1.MESSAGE, "so", "i")+ COUNT(t1.MESSAGE, "some", "i")+
COUNT(t1.MESSAGE, "than", "i")+
COUNT(t1.MESSAGE, "that", "i")+ COUNT(t1.MESSAGE, "the", "i")+
COUNT(t1.MESSAGE, "their", "i")+
COUNT(t1.MESSAGE, "them", "i")+ COUNT(t1.MESSAGE, "then", "i")+
COUNT(t1.MESSAGE, "there", "i")+
COUNT(t1.MESSAGE, "these", "i")+ COUNT(t1.MESSAGE, "they", "i")+
COUNT(t1.MESSAGE, "this", "i")+
COUNT(t1.MESSAGE, "tis", "i")+ COUNT(t1.MESSAGE, "to", "i")+
COUNT(t1.MESSAGE, "too", "i")+
COUNT(t1.MESSAGE, "twas", "i")+ COUNT(t1.MESSAGE, "us", "i")+
COUNT(t1.MESSAGE, "wants", "i")+
COUNT(t1.MESSAGE, "was", "i")+ COUNT(t1.MESSAGE, "we", "i")+
COUNT(t1.MESSAGE, "were", "i")+
COUNT(t1.MESSAGE, "what", "i")+ COUNT(t1.MESSAGE, "when", "i")+
COUNT(t1.MESSAGE, "where", "i")+
COUNT(t1.MESSAGE, "which", "i")+ COUNT(t1.MESSAGE, "while", "i")+
COUNT(t1.MESSAGE, "who", "i")+
COUNT(t1.MESSAGE, "whom", "i")+ COUNT(t1.MESSAGE, "why", "i")+
COUNT(t1.MESSAGE, "will", "i")+
COUNT(t1.MESSAGE, "with", "i")+ COUNT(t1.MESSAGE, "would", "i")+
COUNT(t1.MESSAGE, "yet", "i")+
COUNT(t1.MESSAGE, "you", "i")+ COUNT(t1.MESSAGE, "your", "i"))
LABEL="STOPWORDS" AS STOPWORDS
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;

```

The actual list of english stop words is below.

a,able,about,across,after,all,almost,also,am,among,an,and,any,are,as,at,be,because,been,but,b
y,can,cannot,could,dear,did,do,does,either,else,ever,every,for,from,get,got,has,have,he,her,
hers,him,his,how,however,i,if,in,into,is,it,its,just,least,let,like,likely,may,me,might,most,must,my,n
either,no,nor,not,of,off,often,on,only,or,other,our,own,rather,said,say,says,she,should,since,so,so
me,than,that,the,their,them,then,there,these,they,this,tis,to,too,twas,us,wants,was,we,were,what
,when,where,which,while,who,whom,why,will,with,would,yet,you,your

NON-TEXT Indicators (Emoticons)

Child Sexual Predators overuse Emoticons in their attempts to speak like a child and bond with the victim. Emoticon usages typically decreases with increased age of the on-line poster.

This is also a form of forensic countermeasure. It is fairly easy to go on the internet and find lists of emoticons with conflicting meanings, so if the online predator uses only posts with emoticons, a defense attorney can challenge the meaning of the post in court. Some of the emoticon meanings are context sensitive and may only have meaning because of the message threads in posts before the emoticon. The usage of lesser known emoticons that have dual meanings, combined with “null” posts that put some distance between the previous post and the emoticon can function as a forensic countermeasure.

Emoticon Count or Percentage

Child Sexual Predators overuse Emoticons in their attempts to speak like a child and bond with the victim. Emoticon usages typically decreases with increased age of the on-line poster. Because of this, all of the following indicators can function as variables to detect predators.

1. Percent of Words that are Emoticons
2. Percent of posts that contain Emoticons
3. Percent of posts that are Emoticons only.

Emoticon Obscurity

The 20 most popular emoticons represent 91% of all the emoticons used, and the top 40 represent 99% of the usage. There are currently over 2200 documented emoticons in use around the world. About 400 of these contain asian language characters so they may or may not render correctly in English. That still leaves over 1700 rare emoticons that most people will never see.

During the grooming phase, one of the things that a sexual predator wants to do is appear interesting to a child. One way to do that is to use some obscure emoticons to appear “smart” or “worldly” to the child. By using some rare emoticons that the child has to look up or ask about, the predator increases interest and interaction and furthers the grooming process. The additional benefit to the predator is that many of these rare emoticons may have multiple meanings, and can function as a forensic countermeasure. Predators will often use obscure emoticons along with “null” posts as a distancing tactic to create forensic countermeasures with explicit emoticons and text posts.

We can create an EMO_OBSCURITY score by taking a list of all 1400+ emoticons, counting their frequency of occurrence, converting that to a percent / proportion of total emoticon use and taking an inverse of that value. This creates a power law rating of emoticon rarity, or obscurity, based on actual usage that we can then use to score message posts with. By

summing or averaging these EMO_OBSCURITY scores by user, we create a metric from which we can detect this grooming behaviour.

$$\text{EMO_OBSCURITY} = 1 / P[\text{Emoticon X} \mid \text{Emoticon ALL}]$$

This score will range from about 2.5 for the most popular emoticons to 94 Million for the least used emoticons. Because these emoticons follow a power law distribution of usage, they can become a very strong indicator of predatory behaviour, especially when used in combination with NULL_POST variables. It is important to note that not all sexual predator (are smart enough to) use this approach to grooming or as a countermeasure.

CHIRAL Emoticons

One common form of forensic countermeasure is a Chiral or Left Handed emoticon. Many IM services will render the text of an emoticon into a GIF image or an animated picture. The author knows of no IM service that renders a lefthanded version into pictures. These left handed versions remain as text, but typically cannot be detected or blocked as being obscene. The emoticon for Oral Sex :-* becomes *-:, it conveys the same meaning but will get past ISP blocking, avoids most automatic detection and also becomes a legal forensic countermeasure during court testimony. A count or normalized count of chiral emoticons can be a valuable variable in detecting predatory behaviour.

NULL Posts

Null posts are where a user enters a carriage return, a space, a period (.), a ? or some other placeholder character. In general messaging lingo this means “waiting” or functions as an “are you there?” ping to the other user. Not all IM services support a carriage return as a null post, so the user hits the spacebar and then enter. If the other user is focussing on another browser window, this action will flash the IM icon at the bottom of the screen to get that person’s attention.

It is characteristic of predators to be patient if they are making progress and impatient if they are not.

The use of characters, such as k, K, which are short for “OK” do not count toward a NULL_POST index. Other excludable characters include numbers (0-9), f or F which is short for “fuck” (this can be counted in an explicit word index), u or U which is short for “you” and is often used as a response to a ask the same repeat question or inquiry.

Predator: I am going out tonight.

Predator: u

Victim: Staying home.

Single letters “o” or “O” which mean “Oh” and function as a response statement.

Predator: I am going out tonight.

Victim: o

Single letter “y” or “Y” which means “why” and are a response prompt question.

Victim: I am going out tonight.

Predator: y

The ? symbol may or may not count depending on context. It can also mean a “why” question, but can be a “wait” marker as well.

Victim: I am going out tonight.

Predator:?

or usage as

Predator: Are you there?

Predator: ?

Predator: ?

Predator: ?

Predator: ?

Predator: ?

Predator: ?

Because not all predators use these posts, and the presence of these variations, single character or Null Character posts can be difficult to process. Some require the context of related posts in order to interpret them. Some IM services have active spell check and will not post “Null” or space only posts because they do not appear in a dictionary. These type of services would exclude this type of behavior.

Emotional Sentiment Models

There are a number of Sentiment models based on emotion that can be used in text mining. In this paper we are going to discuss three of those models that are both popular and in general free for use. The main theory behind these models is that different words or emoticons have emotional power and some emotional dimension that relates to the speakers (actors) emotional state. By converting the words to numbers or scores on the different emotional scales we can quantify an emotional state of a speaker, and in turn use that as variables to detect predatory behaviour. These models have found use in many sentiment scoring applications, everything from measuring if customers are happy with a product to detecting emotional stress or depression in addition to predation.

I will cover a brief overview of three emotional models and their use. These models are the

1. the Ekman Emotional Model,
2. Sentiwordnet from Harvard University,
3. Plutchik emotional model.

Along with these models I will also cover the Word Affect scoring model which is more of a word influence model that also has use in detecting predators that are trying to groom a victim.

EKMAN Emotional Model

The Ekman model is based on the work of Paul Ekman and includes the idea of people having 6 basic emotions. These are Anger, Fear, Disgust, Happiness, Sadness and Surprise. Each word or Emoticon is linked to a primary emotion and each word has a polarity and strength. Anger, Fear, Disgust and Sadness have a negative polarity and Happiness and Surprise are positive. The strength scores go either from 0 to 1 or 0 to -1 based on polarity. Some words or emoticons can be neutral and have a score of 0.

Each message can be scanned for words or emoticons and converted to a basic emotion and summed in net strength for a sentiment score. Hence each message post will have six scores and can function as six variables in a statistical model for detecting predators.

The corpus file EMO_Ekman has the emoticon list with the Ekman emotional scores included.

Sentiment Word Score (SentinetWord) [13]

Sentiment word scores works off the idea that all words can be used to convey a sentiment. Sentiment has several dimensions, the first being Positiveness or Negativeness of the word and the second is Objectiveness (or Subjectiveness) of the word. It is possible for a word to convey both positive and negative sentiment at the same time. The scale for SentinetWord becomes 3-axis scale with a polarity net score that is the sum of the positivity score and the negativity score as the first two scales and the Objectivity score as the third scale with is a measure of Subjectivity / Objectivity scale (SO-Polarity)

SO-polarity, as in deciding whether a given text has a factual nature or expresses an opinion on its subject matter. This amounts to performing binary text categorization under categories Subjective and Objective. This is accomplished by the Objective Score of a word

Determining text PN-polarity, as in deciding if a given Subjective text expresses a Positive or a Negative opinion on its subject matter. This happens by the PosScore and NegScore values such that $-1 \leq (\text{PosScore} + \text{NegScore}) \leq 1$.

Positivity and negativity can both be scored as yes=1, no=0 attributes for scoring messages also. This is less sensitive, but computationally easier.

Determining the strength of text PN-polarity, as in deciding e.g. whether the Positive opinion expressed by a text on its subject matter is Weakly Positive, Mildly Positive, or Strongly Positive. Words or emoticons can convey both positive and negative sentiment depending on usage.

Plutchik Emotional Model [10-12]

Robert Plutchik's psychoevolutionary theory of emotion is one of the most influential classification approaches for general emotional responses. He considered there to be eight primary emotions—anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. Plutchik proposed that these 'basic' emotions are biologically primitive and have evolved in order to increase the reproductive fitness of the animal. Plutchik argues for the primacy of these emotions by showing each to be the trigger of behaviour with high survival value, such as the way fear inspires the fight-or-flight response.

Plutchik's psychoevolutionary theory of basic emotions has ten postulates.

1. The concept of emotion is applicable to all evolutionary levels and applies to all animals including humans.
2. Emotions have an evolutionary history and have evolved various forms of expression in different species.
3. Emotions served an adaptive role in helping organisms deal with key survival issues posed by the environment.
4. Despite different forms of expression of emotions in different species, there are certain common elements, or prototype patterns, that can be identified.
5. There is a small number of basic, primary, or prototype emotions.
6. All other emotions are mixed or derivative states; that is, they occur as combinations, mixtures, or compounds of the primary emotions.
7. Primary emotions are hypothetical constructs or idealized states whose properties and characteristics can only be inferred from various kinds of evidence.
8. Primary emotions can be conceptualized in terms of pairs of polar opposites.
9. All emotions vary in their degree of similarity to one another.
10. Each emotion can exist in varying degrees of intensity or levels of arousal.

FORENSIC COUNTERMEASURES

Interrogatives for the basis for questions within the English language and are used by people to request and gather information from others during a conversation. For Sexual predators, this is vital for Step 1 of the Attack Sequence and failure here stops the entire process. Inclusive in this list is the standard ones, who, what, when, where, why, whom, which and how. With the advent of internet slang, a larger list of words, abbreviation and symbols have entered the vernacular that function as interrogatives. This include ?, dig, 3rd degree, 3rd, third degree, fish, grill, hammer, pimp, pump, d, D. The d and D are short for "Details" and get used as follows

VICTIM: I am going out with a friend tonight.

PREDATOR: d

VICTIM: I am going out with a friend tonight.

PREDATOR: D

Words like Fish and Hammer can also have Emoticon substitutes which may also be added to the list at the programmer's discretion.

CURRENT (EXAMPLE) PREDICTION MODEL

For our development, the author was trying to detect predators on a NASCAR fan site sponsored by a company. The site attracted a good percentage of underage fans, but also experienced an above average rate of Explicit word use by the typical, non-predatory fan. This represents a higher than normal rate of noise for the detection model. Any variable that works on the basis of detecting a predators usage of explicit words as being higher than normal, would be reduced in effectiveness (power) as a variable.

Nearly all statistical models will fall into one of 3 categories, 1. Regression models, 2. Neural networks or 3. Decision trees. The author used a regression model for the initial attempts in detecting predators. The basic steps used to develop the model are listed below.

1. Import data, separate into variable fields and clean up data.
2. Calculate and Extract variables that are indicators of sexual predation.
3. Normalize variables against an appropriate base measure (Letters, Words, Posts, etc)
4. Separate data into training and validation data sets.
5. Analyze variables for predictive strength and use statistical profiling to build a model.
6. Prune non-significant variables from model to simplify.
7. Use developed model test against validation set.

In the model development process, only a handful of the 30+ variables were determined to be statistically significant. The author developed several regression models and chose the one that was providing the best correct classification rate. The model and SAS code that was used is below.

```
PROC SQL;
  CREATE TABLE WORK.QUERY_FOR_STATS_WORD_COUNTS_0000 AS
  SELECT t1.NAME,
         t1.SOURCE1,
         t1.Target,
         /* PREDICTION YES=1, or Predator=1*/
         (IFN(ROUND((-0.557) + (29.865 * t1.N_MODALVERBS) + (2.336 *
t1.N_PERPRONOUN) + (-23.643 * t1.N_FAMILY) + (
         t1.N_PERPRONOUN - 0.204) * (( t1.N_REFPRONOUN - 0.001) * 869.195) + (
t1.N_PERPRONOUN - 0.204) * ((
         t1.N_STRETCHWORD - 0.0079) * 179.58),1)<0.5,0,1)
         ) LABEL="PREDICTION" AS PREDICTION,
  FROM WORK.QUERY_FOR_STATS_WORD_COUNTS t1;
QUIT;
```

/* The following code outputs a filtered dataset where the training data set targets=1 or predator and the Prediction values =1 or predator so Type 1 and Type 2 error rates can be estimated */

```
PROC SQL;
  CREATE TABLE WORK.FILTER_FOR_QUERY_FOR_STATS_WORD_ AS
  SELECT t1.NAME,
         t1.SOURCE1,
         t1.Target,
         t1.PREDICTION,
         t1.PREDICTION2
  FROM WORK.QUERY_FOR_STATS_WORD_COUNTS_0000 t1
  WHERE t1.PREDICTION = 1 OR t1.Target = 1;
QUIT;
```

The final model uses only the following variables.

1. Normalized Modal Verb Count
2. Normalized Personal Pronouns
3. Normalized Family Words
4. Normalized Stretch Word usage

Variables like Explicit Words or Reflexive Personal Pronouns dropped out of the model. It is very likely that a sexual predators use of explicit words affected these variables even if the direct

usage of those variables did not happen in this model. A number of other variables did not make the cutoff for being statistically significant, but the author suspects that on a website where the average usage of explicit words is lower, these variables would become stronger indicators and likely get included in a model. FIGURE 4 below has some regression graphs on some of the key variables in the model.

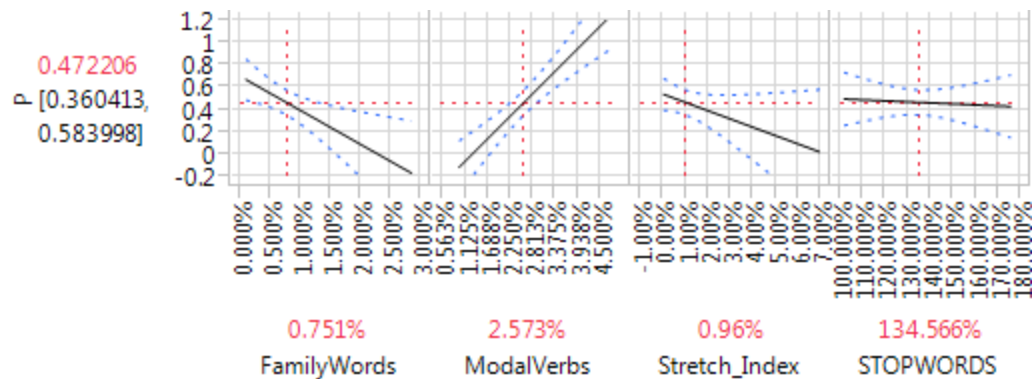


FIGURE 4: Regression graphs some key variables of the authors model.

OTHER ANALYSIS AND REPORTING

The real final end product that was desired from this system was a daily output report to the company security team, that showed the Username, Predator prediction and which social media feed the user was on, so that security personnel could do a manual review of the posts and then block anyone confirmed as a likely predator. An example of the output table is below along with the code that generates it.

Summary Tables		
		PREDICTION
NAME		
	N	1
alonzo403	N	1
andaconda40	N	1
artc68	N	1
audavideo	N	1
augustadude2004	N	1
bald_bever_hunter42025	N	1
balin79	N	1
bayouboy924	N	1
bearinwolfsfur	N	1
bendix632	N	1
big_dave5188	N	1
bushsucksgrrl	N	1
can_i_rape_you_anally	N	1
chrisnr1998	N	1
cse1606	N	1
evyfresh93	N	1
georgia_emi	N	1
hhhhgoatsrock	N	1
jackies2cool4u	N	1
jaguarsd69	N	1
jeff_fisher94	N	1
jon_raven2000	N	1
kckboxer18	N	1
kevin	N	1

FIGURE 5: Sample daily output report for the security team.

```

PROC TABULATE
DATA=WORK.FILTER_FOR_QUERY_FOR_STATS_WORD_;
    VAR PREDICTION PREDICTION2;
    CLASS NAME /      ORDER=UNFORMATTED MISSING;
    CLASS SOURCE1 /   ORDER=UNFORMATTED MISSING;
    CLASS Target /    ORDER=UNFORMATTED MISSING;
    TABLE /* Row Dimension */
SOURCE1*
    NAME*
    N,
/* Column Dimension */
PREDICTION
PREDICTION2 ;
    ;
RUN; QUIT;
TITLE; FOOTNOTE;

```

REFERENCES

1. Lanning, KV. "Child molesters: A behavioral analysis." 2001. Federal Bureau of Investigation.
2. McGhee, India et al. "Learning to identify Internet sexual predation." International Journal of Electronic Commerce 15.3 (2011): 103-122.
3. <http://www.datagenetics.com/blog/october52012/index.html>
4. Exploiting Emoticons in Sentiment Analysis,
<http://people.few.eur.nl/frasincar/papers/SAC2013b/sac2013b.pdf> Alexander Hogenboom1 hogenboom@ese.eur.nl Daniella Bal1 daniella.bal@xs4all.nl Flavius Frasinicar1 frasincar@ese.eur.nl Malissa Bal1 malissa.bal@xs4all.nl Franciska de Jong1,2 f.m.g.dejong@utwente.nl Uzay Kaymak3 u.kaymak@ieee.org
5. <http://www.umiacs.umd.edu/~saif/WebPages/Abstracts/NRC-SentimentAnalysis.htm>
6. McMillan dictionary online - relation words.
<http://www.macmillandictionary.com/us/thesaurus-category/american/words-used-to-describe-relations-and-relationships>
7. <http://www.thesaurus.com/browse/apology> Apology word corpus source list.
8. http://pillsbury.mpls.k12.mn.us/uploads/pronouns_in_first_second_third_person.pdf
Personal Pronoun list
9. Child Molesters:a behavioural analysis. Kenneth V. Lanning, Supervisory Special Agent Behavioral Science Unit, FBI, Quantico VA. 1992, Center for Missing and Exploited Children.
10. Plutchik, Robert (1980), Emotion: Theory, research, and experience: Vol. 1. Theories of emotion 1, New York: Academic
11. Plutchik, Robert (2002), Emotions and Life: Perspectives from Psychology, Biology, and Evolution, Washington, DC: American Psychological Association
12. Plutchik, Robert; R. Conte., Hope (1997), Circumplex Models of Personality and Emotions, Washington, DC: American Psychological Association
13. Sentiwordnet sentiment corpus project. <http://sentiwordnet.isti.cnr.it/>
14. <https://github.com/ahaque/twitch-troll-detection>
15. Bogdanova, Rosso, Solorio, (2012), On the impact of sentiment and Emotion Based Features in Detecting Online Sexual Predators. Assoc. for Computational Linguistics.
16. Peersman, Vaassen, Van Asch, Daclemans, (2012) Conversation Level Constraints on Pedophile Detection in Chat Rooms. Antwerp University.
17. Tschuggnall, Specht, (2014), What Grammar Tells about Gender and Age of Authors. University of Innsbruck, Austria.

CORPUS APPENDIXES

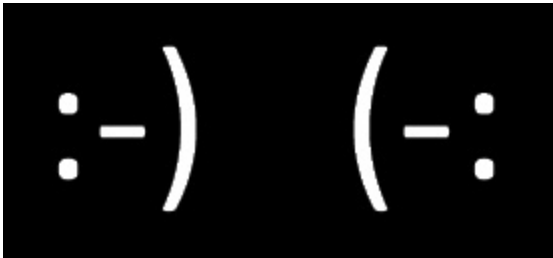
APPENDIX A: The top emoticons [3]

It's no big surprise, but the popular emoticons dominate the usage patterns. Of the **96,269,892** Tweets that contained emoticons, the top 20 smileys accounted for 90% of all occurrences.

	Emoticon	Usage	Percent	Notes
#1	:)	32,115,789	33.360%	<i>Happy face</i>
#2	:D	10,595,385	11.006%	<i>Laugh</i>
#3	:(7,613,014	7.908%	<i>Sad face</i>
#4	;)	7,238,295	7.519%	<i>Wink</i>
#5	: -)	4,254,708	4.420%	<i>Happy face (with nose)</i>
#6	:P	3,588,863	3.728%	<i>Tongue out</i>
#7	=)	3,564,080	3.702%	<i>Happy face</i>
#8	(:	2,720,383	2.826%	<i>Happy face (mirror)</i>
#9	;-)	2,085,015	2.166%	<i>Wink (with nose)</i>
#10	:/	1,840,827	1.912%	<i>Uneasy, undecided, skeptical, annoyed?</i>
#11	XD	1,795,792	1.865%	<i>Big grin</i>
#12	=D	1,434,004	1.490%	<i>Laugh</i>
#13	:o	1,077,124	1.119%	<i>Shock, Yawn</i>
#14	=]	1,055,517	1.096%	<i>Happy face</i>
#15	D:	1,048,320	1.089%	<i>Grin (mirror)</i>
#16	;D	1,004,509	1.043%	<i>Wink and grin</i>
#17	:]	954,740	0.992%	<i>Happy face</i>
#18	:-(816,170	0.848%	<i>Unhappy</i>
#19	=/	809,760	0.841%	<i>Uneasy, undecided, skeptical, annoyed?</i>

#20	=(760,600	0.790%	<i>Unhappy</i>
-----	----	---------	--------	----------------

Chiral versions of most emoticons (where the mouth occurs before the eyes, and not the other way around) occur with a ratio of approximately 1:12. (I have no data, nor performed research into this, but coincidentally, this ratio is approximately the same as ratio of *right-handed* to *left-handed* people).



As an example, imagine the simple phrase *"I'm going to bed early tonight"*.

I'm going to bed early tonight :-)	Happy smile
I'm going to bed early tonight :-(Sad frown
I'm going to bed early tonight :-	Angry
I'm going to bed early tonight ;-)	Wink
I'm going to bed early tonight :-D	Deep laugh
I'm going to bed early tonight :-/	Uneasy
I'm going to bed early tonight :-P	Tongue out

APPENDIX B: Desensitising words or Explicit Word Corpus.

anal
anus
arse
ass
ballsack
balls

bastard
bitch
biatch
bloody
blowjob
blow job
bollock
bollok
boner
boob
bugger
bum
butt
buttplug
clitoris
cock
coon
crap
cunt
damn
dick
dildo
dyke
fag
feck
fellate
fellatio
felching
fuck
f u c k
fudgepacker
fudge packer
flange
Goddamn
God damn
hell
homo
jerk
jizz
knobend
knob end
labia
Imao

lmfao
muff
nigger
nigga
omg
penis
piss
poop
prick
pube
pussy
queer
scrotum
sex
shit
s hit
sh1t
slut
smegma
spunk
tit
tosser
turd
twat
vagina
wank
whore
wtf

APPENDIX C: Apology Words

[7] <http://www.thesaurus.com/browse/apology>

apologise
apologize
apologies
sorry
sorrry
soorry
sorree
sorrii
me bad

my bad
mea bad
i bad
Unfortunately
I'm afraid that
bad news
puts you out
put you out
acknowledgment
admission
amend
atonement
concession
confession
excuse
explanation
justification
justified
plea
redress
reparation
defense
extenuation
extenuating
mitigation
mitigate
vindication
vindicate
mea culpa
denial
refusal
refuse
repudiation
repudiate
accusation
accuse
censure
charge
complaint
condemnation
criticism
defiance
indictment

apology
atonement
expiation
indemnification
indemnity
quittance
recompense
redress
reparation
reprisal
requital
restitution
restoration
satisfaction
satisfy
answer
apologetics
apologia
apologizing
apology
argument
cleanup
copout
exculpation
excuse
excusing
exoneration
explaining
extenuation
fish story
jive
off-time
plea
rationalization
rejoinder
reply
response
retort
return
song and dance
story
vindication
whitewash

alibi
apology
cleanup
cop-out
cover
cover story
coverup
defense
disguise
evasion
expedient
extenuation
fish story
grounds
jive
justification
makeshift
mitigation
plea
pretext
rationalization
regrets
routine
semblance
shift
song
song and dance
stall
stopgap
story
substitute
subterfuge
trick
vindication
whitewash
why and wherefore
wherefore and why
why
wherefore
absolution
account
acquittal
advocacy

answer
apologia
apology
approval
argument
basis
confirmation
defense
exculpation
exoneration
explanation
extenuation
grounds
idea
mitigation
palliation
palliative
plea
pretext
raison d'être
rationale
rationalization
rebuttal
redemption
reply
response
salvation
sanctification
song and dance
story
support
validation
vindication
warrant
whatfor
wherefore
whitewashing
whole idea
apology
burlesque
caricature
cartoon
copy

derision
farce
irony
jest
joke
lampoon
mime
mimicry
misrepresentation
misrepresent
mock-heroic
mockery
pastiche
play-on
raillery
rib
ridicule
roast
satire
send-up
skit
takeoff
travesty

APPENDIX D: PERSONAL PRONOUN CORPUS

[8]

http://pillsbury.mpls.k12.mn.us/uploads/pronouns_in_first_second_third_person.pdf

PRONOUN	PERSON	CASE	PLURALITY
I	FIRST	SUBJECTIVE	SINGULAR
we	FIRST	SUBJECTIVE	PLURAL
me	FIRST	OBJECTIVE	SINGULAR
us	FIRST	OBJECTIVE	PLURAL
my	FIRST	POSSESSIVE	SINGULAR
our	FIRST	POSSESSIVE	PLURAL
mine	FIRST	POSSESSIVE	SINGULAR

ours	FIRST	POSSESSIVE	PLURAL
you	SECOND	SUB/OBJ	BOTH
your	SECOND	POSSESSIVE	SINGULAR
yours	SECOND	POSSESSIVE	PLURAL
he	THIRD	SUBJECTIVE	SINGULAR
she	THIRD	SUBJECTIVE	SINGULAR
it	THIRD	SUBJECTIVE	SINGULAR
they	THIRD	SUBJECTIVE	PLURAL
him	THIRD	OBJECTIVE	SINGULAR
her	THIRD	OBJECTIVE	SINGULAR
it	THIRD	OBJECTIVE	SINGULAR
them	THIRD	OBJECTIVE	PLURAL
his	THIRD	POSSESSIVE	SINGULAR
her	THIRD	POSSESSIVE	SINGULAR
hers	THIRD	POSSESSIVE	SINGULAR
its	THIRD	POSSESSIVE	SINGULAR
their	THIRD	POSSESSIVE	PLURAL
theirs	THIRD	POSSESSIVE	PLURAL

Appendix E: Interrogative Word Corpus

who
 hwo
 what
 when
 where
 why
 how
 whom
 which
 ?

dig
3rd degree
3rd
third degree
fish
grill
hammer
pimp
pump
d
D

SAS CODE APPENDIXES

Appendix C1: DESENSWORD INDEX CODE

```
PROC SQL;  
  CREATE TABLE  
WORK.QUERY_FOR_CHATLOG_0001(label="WORK.QUERY_FOR_CHATLOG_0001") AS  
  SELECT t1.F1,  
    t1.NAME,  
    t1.DATETIME,  
    t1.DATE,  
    t1.TIME,  
    t1.MESSAGE,  
    /* DESENSWORD */  
    (COUNT(t1.MESSAGE, "anal", "i")  
    + COUNT(t1.MESSAGE, "anus", "i")  
    + COUNT(t1.MESSAGE, "arse", "i")  
    + COUNT(t1.MESSAGE, "ass", "i")  
    + COUNT(t1.MESSAGE, "ballsack", "i")  
    + COUNT(t1.MESSAGE, "balls", "i")  
    + COUNT(t1.MESSAGE, "bastard", "i")  
    + COUNT(t1.MESSAGE, "bitch", "i")  
    + COUNT(t1.MESSAGE, "biatch", "i")  
    + COUNT(t1.MESSAGE, "bloody", "i")  
    + COUNT(t1.MESSAGE, "blowjob", "i")  
    + COUNT(t1.MESSAGE, "blow job", "i")  
    + COUNT(t1.MESSAGE, "bollock", "i")  
    + COUNT(t1.MESSAGE, "bollok", "i")  
    + COUNT(t1.MESSAGE, "boner", "i"))
```

+ COUNT(t1.MESSAGE, "boob", "i")
+ COUNT(t1.MESSAGE, "bugger", "i")
+ COUNT(t1.MESSAGE, "bum", "i")
+ COUNT(t1.MESSAGE, "butt", "i")
+ COUNT(t1.MESSAGE, "buttplug", "i")
+ COUNT(t1.MESSAGE, "clitoris", "i")
+ COUNT(t1.MESSAGE, "cock", "i")
+ COUNT(t1.MESSAGE, "coon", "i")
+ COUNT(t1.MESSAGE, "crap", "i")
+ COUNT(t1.MESSAGE, "cunt", "i")
+ COUNT(t1.MESSAGE, "damn", "i")
+ COUNT(t1.MESSAGE, "dick", "i")
+ COUNT(t1.MESSAGE, "dildo", "i")
+ COUNT(t1.MESSAGE, "dyke", "i")
+ COUNT(t1.MESSAGE, "fag", "i")
+ COUNT(t1.MESSAGE, "feck", "i")
+ COUNT(t1.MESSAGE, "fellate", "i")
+ COUNT(t1.MESSAGE, "fellatio", "i")
+ COUNT(t1.MESSAGE, "felching", "i")
+ COUNT(t1.MESSAGE, "fuck", "i")
+ COUNT(t1.MESSAGE, "f u c k", "i")
+ COUNT(t1.MESSAGE, "fudgepacker", "i")
+ COUNT(t1.MESSAGE, "fudge packer", "i")
+ COUNT(t1.MESSAGE, "flange", "i")
+ COUNT(t1.MESSAGE, "Goddamn", "i")
+ COUNT(t1.MESSAGE, "God damn", "i")
+ COUNT(t1.MESSAGE, "hell", "i")
+ COUNT(t1.MESSAGE, "homo", "i")
+ COUNT(t1.MESSAGE, "jerk", "i")
+ COUNT(t1.MESSAGE, "jizz", "i")
+ COUNT(t1.MESSAGE, "knobend", "i")
+ COUNT(t1.MESSAGE, "knob end", "i")
+ COUNT(t1.MESSAGE, "labia", "i")
+ COUNT(t1.MESSAGE, "lmao", "i")
+ COUNT(t1.MESSAGE, "lmfao", "i")
+ COUNT(t1.MESSAGE, "muff", "i")
+ COUNT(t1.MESSAGE, "nigger", "i")
+ COUNT(t1.MESSAGE, "nigga", "i")
+ COUNT(t1.MESSAGE, "omg", "i")
+ COUNT(t1.MESSAGE, "penis", "i")
+ COUNT(t1.MESSAGE, "piss", "i")
+ COUNT(t1.MESSAGE, "poop", "i")
+ COUNT(t1.MESSAGE, "prick", "i")

```

+ COUNT(t1.MESSAGE, "pube", "i")
+ COUNT(t1.MESSAGE, "pussy", "i")
+ COUNT(t1.MESSAGE, "queer", "i")
+ COUNT(t1.MESSAGE, "scrotum", "i")
+ COUNT(t1.MESSAGE, "sex", "i")
+ COUNT(t1.MESSAGE, "shit", "i")
+ COUNT(t1.MESSAGE, "s hit", "i")
+ COUNT(t1.MESSAGE, "sh1t", "i")
+ COUNT(t1.MESSAGE, "slut", "i")
+ COUNT(t1.MESSAGE, "smegma", "i")
+ COUNT(t1.MESSAGE, "spunk", "i")
+ COUNT(t1.MESSAGE, "tit", "i")
+ COUNT(t1.MESSAGE, "tosser", "i")
+ COUNT(t1.MESSAGE, "turd", "i")
+ COUNT(t1.MESSAGE, "twat", "i")
+ COUNT(t1.MESSAGE, "vagina", "i")
+ COUNT(t1.MESSAGE, "wank", "i")
+ COUNT(t1.MESSAGE, "whore", "i")
+ COUNT(t1.MESSAGE, "wtf", "i")) LABEL="DESENSWORD" AS DESENSWORD
FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;

```

Appendix C2: APOLOGYWORD Index Code

```

PROC SQL;
  CREATE TABLE
WORK.QUERY_FOR_CHATLOG_0001(label="WORK.QUERY_FOR_CHATLOG_0001") AS
  SELECT t1.F1,
    t1.NAME,
    t1.DATETIME,
    t1.DATE,
    t1.TIME,
    t1.MESSAGE,
    /* APOLOGYWORD */
    (COUNT(t1.MESSAGE, "apologise", "i")
+ COUNT(t1.MESSAGE, "apologize", "i")
+ COUNT(t1.MESSAGE, "apologies", "i")
+ COUNT(t1.MESSAGE, "sorry", "i")
+ COUNT(t1.MESSAGE, "sorrry", "i")
+ COUNT(t1.MESSAGE, "soorry", "i")
+ COUNT(t1.MESSAGE, "sorree", "i")
+ COUNT(t1.MESSAGE, "sorrii", "i")
+ COUNT(t1.MESSAGE, "me bad", "i"))

```

+ COUNT(t1.MESSAGE, "my bad", "i")
+ COUNT(t1.MESSAGE, "mea bad", "i")
+ COUNT(t1.MESSAGE, "i bad", "i")
+ COUNT(t1.MESSAGE, "Unfortunately", "i")
+ COUNT(t1.MESSAGE, "I'm afraid that", "i")
+ COUNT(t1.MESSAGE, "bad news", "i")
+ COUNT(t1.MESSAGE, "puts you out", "i")
+ COUNT(t1.MESSAGE, "put you out", "i")
+ COUNT(t1.MESSAGE, "acknowledgment", "i")
+ COUNT(t1.MESSAGE, "admission", "i")
+ COUNT(t1.MESSAGE, "amend", "i")
+ COUNT(t1.MESSAGE, "atonement", "i")
+ COUNT(t1.MESSAGE, "concession", "i")
+ COUNT(t1.MESSAGE, "confession", "i")
+ COUNT(t1.MESSAGE, "excuse", "i")
+ COUNT(t1.MESSAGE, "explanation", "i")
+ COUNT(t1.MESSAGE, "justification", "i")
+ COUNT(t1.MESSAGE, "justified", "i")
+ COUNT(t1.MESSAGE, "plea", "i")
+ COUNT(t1.MESSAGE, "redress", "i")
+ COUNT(t1.MESSAGE, "reparation", "i")
+ COUNT(t1.MESSAGE, "defense", "i")
+ COUNT(t1.MESSAGE, "extenuation", "i")
+ COUNT(t1.MESSAGE, "extenuating", "i")
+ COUNT(t1.MESSAGE, "mitigation", "i")
+ COUNT(t1.MESSAGE, "mitigate", "i")
+ COUNT(t1.MESSAGE, "vindication", "i")
+ COUNT(t1.MESSAGE, "vidicate", "i")
+ COUNT(t1.MESSAGE, "mea culpa", "i")
+ COUNT(t1.MESSAGE, "denial", "i")
+ COUNT(t1.MESSAGE, "refusal", "i")
+ COUNT(t1.MESSAGE, "refuse", "i")
+ COUNT(t1.MESSAGE, "repudiation", "i")
+ COUNT(t1.MESSAGE, "repudiate", "i")
+ COUNT(t1.MESSAGE, "accusation", "i")
+ COUNT(t1.MESSAGE, "accuse", "i")
+ COUNT(t1.MESSAGE, "censure", "i")
+ COUNT(t1.MESSAGE, "charge", "i")
+ COUNT(t1.MESSAGE, "complaint", "i")
+ COUNT(t1.MESSAGE, "condemnation", "i")
+ COUNT(t1.MESSAGE, "criticism", "i")
+ COUNT(t1.MESSAGE, "defiance", "i")
+ COUNT(t1.MESSAGE, "indictment", "i")
+ COUNT(t1.MESSAGE, "apology", "i")
+ COUNT(t1.MESSAGE, "atonement", "i")

+ COUNT(t1.MESSAGE, "expiation", "i")
+ COUNT(t1.MESSAGE, "indemnification", "i")
+ COUNT(t1.MESSAGE, "indemnity", "i")
+ COUNT(t1.MESSAGE, "quittance", "i")
+ COUNT(t1.MESSAGE, "recompense", "i")
+ COUNT(t1.MESSAGE, "redress", "i")
+ COUNT(t1.MESSAGE, "reparation", "i")
+ COUNT(t1.MESSAGE, "reprisal", "i")
+ COUNT(t1.MESSAGE, "requital", "i")
+ COUNT(t1.MESSAGE, "restitution", "i")
+ COUNT(t1.MESSAGE, "restoration", "i")
+ COUNT(t1.MESSAGE, "satisfaction", "i")
+ COUNT(t1.MESSAGE, "satisfy", "i")
+ COUNT(t1.MESSAGE, "answer", "i")
+ COUNT(t1.MESSAGE, "apologetics", "i")
+ COUNT(t1.MESSAGE, "apologia", "i")
+ COUNT(t1.MESSAGE, "apologizing", "i")
+ COUNT(t1.MESSAGE, "apology", "i")
+ COUNT(t1.MESSAGE, "argument", "i")
+ COUNT(t1.MESSAGE, "cleanup", "i")
+ COUNT(t1.MESSAGE, "copout", "i")
+ COUNT(t1.MESSAGE, "exculpation", "i")
+ COUNT(t1.MESSAGE, "excuse", "i")
+ COUNT(t1.MESSAGE, "excusing", "i")
+ COUNT(t1.MESSAGE, "exoneration", "i")
+ COUNT(t1.MESSAGE, "explaining", "i")
+ COUNT(t1.MESSAGE, "extenuation", "i")
+ COUNT(t1.MESSAGE, "fish story", "i")
+ COUNT(t1.MESSAGE, "jive", "i")
+ COUNT(t1.MESSAGE, "off-time", "i")
+ COUNT(t1.MESSAGE, "plea", "i")
+ COUNT(t1.MESSAGE, "rationalization", "i")
+ COUNT(t1.MESSAGE, "rejoinder", "i")
+ COUNT(t1.MESSAGE, "reply", "i")
+ COUNT(t1.MESSAGE, "response", "i")
+ COUNT(t1.MESSAGE, "retort", "i")
+ COUNT(t1.MESSAGE, "return", "i")
+ COUNT(t1.MESSAGE, "song and dance", "i")
+ COUNT(t1.MESSAGE, "story", "i")
+ COUNT(t1.MESSAGE, "vindication", "i")
+ COUNT(t1.MESSAGE, "whitewash", "i")
+ COUNT(t1.MESSAGE, "alibi", "i")
+ COUNT(t1.MESSAGE, "apology", "i")
+ COUNT(t1.MESSAGE, "cleanup", "i")
+ COUNT(t1.MESSAGE, "cop-out", "i")

+ COUNT(t1.MESSAGE, "cover", "i")
+ COUNT(t1.MESSAGE, "cover story", "i")
+ COUNT(t1.MESSAGE, "coverup", "i")
+ COUNT(t1.MESSAGE, "defense", "i")
+ COUNT(t1.MESSAGE, "disguise", "i")
+ COUNT(t1.MESSAGE, "evasion", "i")
+ COUNT(t1.MESSAGE, "expedient", "i")
+ COUNT(t1.MESSAGE, "extenuation", "i")
+ COUNT(t1.MESSAGE, "fish story", "i")
+ COUNT(t1.MESSAGE, "grounds", "i")
+ COUNT(t1.MESSAGE, "jive", "i")
+ COUNT(t1.MESSAGE, "justification", "i")
+ COUNT(t1.MESSAGE, "makeshift", "i")
+ COUNT(t1.MESSAGE, "mitigation", "i")
+ COUNT(t1.MESSAGE, "plea", "i")
+ COUNT(t1.MESSAGE, "pretext", "i")
+ COUNT(t1.MESSAGE, "rationalization", "i")
+ COUNT(t1.MESSAGE, "regrets", "i")
+ COUNT(t1.MESSAGE, "routine", "i")
+ COUNT(t1.MESSAGE, "semblance", "i")
+ COUNT(t1.MESSAGE, "shift", "i")
+ COUNT(t1.MESSAGE, "song", "i")
+ COUNT(t1.MESSAGE, "song and dance", "i")
+ COUNT(t1.MESSAGE, "stall", "i")
+ COUNT(t1.MESSAGE, "stopgap", "i")
+ COUNT(t1.MESSAGE, "story", "i")
+ COUNT(t1.MESSAGE, "substitute", "i")
+ COUNT(t1.MESSAGE, "subterfuge", "i")
+ COUNT(t1.MESSAGE, "trick", "i")
+ COUNT(t1.MESSAGE, "vindication", "i")
+ COUNT(t1.MESSAGE, "whitewash", "i")
+ COUNT(t1.MESSAGE, "why and wherefore", "i")
+ COUNT(t1.MESSAGE, "wherefore and why", "i")
+ COUNT(t1.MESSAGE, "why", "i")
+ COUNT(t1.MESSAGE, "wherefore", "i")
+ COUNT(t1.MESSAGE, "absolution", "i")
+ COUNT(t1.MESSAGE, "account", "i")
+ COUNT(t1.MESSAGE, "acquittal", "i")
+ COUNT(t1.MESSAGE, "advocacy", "i")
+ COUNT(t1.MESSAGE, "answer", "i")
+ COUNT(t1.MESSAGE, "apologia", "i")
+ COUNT(t1.MESSAGE, "apology", "i")
+ COUNT(t1.MESSAGE, "approval", "i")
+ COUNT(t1.MESSAGE, "argument", "i")
+ COUNT(t1.MESSAGE, "basis", "i")

+ COUNT(t1.MESSAGE, "confirmation", "i")
+ COUNT(t1.MESSAGE, "defense", "i")
+ COUNT(t1.MESSAGE, "exculpation", "i")
+ COUNT(t1.MESSAGE, "exoneration", "i")
+ COUNT(t1.MESSAGE, "explanation", "i")
+ COUNT(t1.MESSAGE, "extenuation", "i")
+ COUNT(t1.MESSAGE, "grounds", "i")
+ COUNT(t1.MESSAGE, "idea", "i")
+ COUNT(t1.MESSAGE, "mitigation", "i")
+ COUNT(t1.MESSAGE, "palliation", "i")
+ COUNT(t1.MESSAGE, "palliative", "i")
+ COUNT(t1.MESSAGE, "plea", "i")
+ COUNT(t1.MESSAGE, "pretext", "i")
+ COUNT(t1.MESSAGE, "raison d'être", "i")
+ COUNT(t1.MESSAGE, "rationale", "i")
+ COUNT(t1.MESSAGE, "rationalization", "i")
+ COUNT(t1.MESSAGE, "rebuttal", "i")
+ COUNT(t1.MESSAGE, "redemption", "i")
+ COUNT(t1.MESSAGE, "reply", "i")
+ COUNT(t1.MESSAGE, "response", "i")
+ COUNT(t1.MESSAGE, "salvation", "i")
+ COUNT(t1.MESSAGE, "sanctification", "i")
+ COUNT(t1.MESSAGE, "song and dance", "i")
+ COUNT(t1.MESSAGE, "story", "i")
+ COUNT(t1.MESSAGE, "support", "i")
+ COUNT(t1.MESSAGE, "validation", "i")
+ COUNT(t1.MESSAGE, "vindication", "i")
+ COUNT(t1.MESSAGE, "warrant", "i")
+ COUNT(t1.MESSAGE, "whatfor", "i")
+ COUNT(t1.MESSAGE, "wherefore", "i")
+ COUNT(t1.MESSAGE, "whitewashing", "i")
+ COUNT(t1.MESSAGE, "whole idea", "i")
+ COUNT(t1.MESSAGE, "apology", "i")
+ COUNT(t1.MESSAGE, "burlesque", "i")
+ COUNT(t1.MESSAGE, "caricature", "i")
+ COUNT(t1.MESSAGE, "cartoon", "i")
+ COUNT(t1.MESSAGE, "copy", "i")
+ COUNT(t1.MESSAGE, "derision", "i")
+ COUNT(t1.MESSAGE, "farce", "i")
+ COUNT(t1.MESSAGE, "irony", "i")
+ COUNT(t1.MESSAGE, "jest", "i")
+ COUNT(t1.MESSAGE, "joke", "i")
+ COUNT(t1.MESSAGE, "lampoon", "i")
+ COUNT(t1.MESSAGE, "mime", "i")
+ COUNT(t1.MESSAGE, "mimicry", "i")

```
+ COUNT(t1.MESSAGE, "misrepresentation", "i")
+ COUNT(t1.MESSAGE, "misrepresent", "i")
+ COUNT(t1.MESSAGE, "mock-heroic", "i")
+ COUNT(t1.MESSAGE, "mockery", "i")
+ COUNT(t1.MESSAGE, "pastiche", "i")
+ COUNT(t1.MESSAGE, "play-on", "i")
+ COUNT(t1.MESSAGE, "raillery", "i")
+ COUNT(t1.MESSAGE, "rib", "i")
+ COUNT(t1.MESSAGE, "ridicule", "i")
+ COUNT(t1.MESSAGE, "roast", "i")
+ COUNT(t1.MESSAGE, "satire", "i")
+ COUNT(t1.MESSAGE, "send-up", "i")
+ COUNT(t1.MESSAGE, "skit", "i")
+ COUNT(t1.MESSAGE, "takeoff", "i")
+ COUNT(t1.MESSAGE, "travesty", "i")) LABEL="APOLOGYWORD" AS
APOLOGYWORD
  FROM WORK.QUERY_FOR_CHATLOG_0000 t1;
QUIT;
```