Strategies for neutralising sexually explicit language

George R. S. Weir and Ana-Maria Duta
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
e-mail: george.weir@strath.ac.uk

Abstract*— While scales for the 'strength' of pornographic images have been developed and used in research and for judicial purposes, the potential emotive impact of text-based pornography has received little attention. In this paper we describe our approach to characterising dimensions of sexually explicit language and outline their use in strategies for neutralising such language. Our intended application of this work is contexts where we may wish to affect the content to which a student, trainee or other professional may be exposed. By controlling the quantity and degree of such content, we aim to minimise any detrimental effects (observer impact) that such content may have on ill-prepared individuals.

***Keywords-** sexually explicit text; dimensions of sexual language; neutralising emotive force*

I. Introduction

Exposure to hard-core pornography in the form of explicit images and video is often a necessary aspect of the professional's role in digital forensics, law enforcement or legal representation. In presenting court evidence that includes paedophile imagery, a common strategy is to register the 'strength' of each item as a shorthand record of the content and as a means of conveying the seriousness of any offence. In Europe, the COPINE (Combating Paedophile Information Networks in Europe) classification model is often used [1]. This model, conceived as an aid both to research and criminal proceedings, categorises the severity of victimisation in child pornography on a ten point scale [2]. Based on the COPINE classification, the Sentencing Advisor Panel of the UK's judiciary developed the five point SAP scale for use in court cases [3].

Clearly, a primary motivation behind such an approach is the need to provide a standard for gauging the seriousness of offences. A further motivation is a desire to limit the need to expose individuals to potentially distressing or damaging materials. Thereby, an outline description of the classification scheme, added to a COPINE or SAP measure of the materials in a suspect's possession, can convey one significant dimension in a criminal case, i.e., the degree or 'strength' of any individual image or video content. (A second significant dimension will be the quantity of materials that fall under each classification.)

One major area of concern for law enforcement is the use of social networking and similar chat-based sites for the sexual grooming of minors. According to the Child Exploitation and Online Protection (CEOP) Centre,

> Whilst children can make themselves vulnerable in relation to their online behaviour it is equally the case that offenders target and exploit this vulnerability. CEOP sees frequent grooming behaviour targeted towards children and social networking sites are the most commonly reported environment in which this activity takes place. This is likely to be due to the ease with which individuals can create profiles on such sites. [4, p.7]

Through such networking sites, criminals may seek to develop a relationship with one or more minors as part of the grooming process toward child exploitation.

> Criminals use a number of methods to build friendships with children on social networking sites. Once a relationship is formed the offender will use their position to initiate some form of sexually exploitative activity". (op. cit.)

In such settings, the content may be entirely (or largely) text-based and the collation of evidence against an offender will require the assembly and annotation of such computer-based interactions. As in the graphic image setting noted earlier, we have several contexts in which exposure to explicit materials may be a professional requirement - as a digital forensic investigator, a legal professional or a member of law enforcement.

In the present paper, our application domain is the range of professional contexts in which individuals may be exposed to sexually explicit content, but content that is text-based rather than image-based. Our purpose is to explore the possibility of developing techniques, based upon the classification of such textual content, which will provide means to neutralise by varying degrees, the content to which a student, trainee or other professional may be exposed. This purpose is motivated by a desire to control the quantity and degree of such content in the hope of minimising any detrimental effects (observer impact) that such content may have on ill-prepared individuals. Given that such raw material may be highly sexualised and explicit, we see such a development as a component in the duty of care toward professionals early in a career who may ultimately require first-hand exposure to sexually explicit or paedophilic materials.

II. An Easy Solution?

If our objective was purely evidential, then there would appear to be an easy solution to the classification of such text-based materials. Having derived the content from a suspect's on-line interactions with a child, we might apply the existing COPINE or SAP classification. Let us consider this prospect with reference to the SAP categories [5] listed in Table 1.

TABLE I. SAP IMAGE CLASSIFICATION

1.	Images depicting nudity or erotic posing with no sexual activity
2.	Sexual activity between children, or solo masturbation by a child
3.	Non-penetrative sexual activity between adult(s) and child(ren)
4.	Penetrative sexual activity between child(ren) and adult(s)
5.	Sadism or bestiality

These five categories of paedophile images present a clear ranking of content severity and, if we ignore the explicit reference to images in the first category, the descriptors could be applied to texts. For instance, a paedophile may have a blog that records exploits with children and such activity as described in the blog may map directly to one or more categories in the SAP classification. While such text contexts may allow the evidential use of the SAP categories, one characteristic of this classification is the **presumption of reality**.

When dealing with images, there is reasonable assurance that they represent real events that have actually taken place. Indeed, as a step toward prevention of further exploitation, law enforcement agencies exert considerable energy in attempting to identify children from their appearance in paedophile images. So, the evidential application of such image classification schemes relies upon this assumption that the content so described represents real events. Another way of expressing this point is that the primary evidential basis of such imagery is its role as proof that offences have been committed. The SAP categories can equally apply to 'fabricated' images in which no real children are represented or exploited. In such cases, possession of these images may serve a secondary evidential basis toward establishing the character or inclination of a suspect, but since they would not be tied to any real event, they are not evidence of child exploitation.

The presumption of reality is less strong when dealing with textual descriptions of events. This is why we often have to suspend disbelief when reading novels or listening to radio plays. Of course, textual accounts (like verbal accounts) may provide corroborative evidence that can be tied to other sources and known events, e.g., through the presence of precise details that are unlikely to be known by non-participants or non-observers to the actual event. Nevertheless, the prima facie evidence that an event has taken place is less strong in the case of a textual account than when presented by a photographic record.

A further significant mismatch between the SAP-style image classification and text-based content lies in the nature of offence for which the content provides evidence. In the case of imagery, the possessor may be guilty of an offence in virtue of possession or distribution of such content. This is unlikely to be the case for textual materials. (Of course, this will depend upon the prevailing detail of the obscene publications act.) The offence more commonly committed in a text-based child interaction setting is that of grooming. Significantly, images have no required role in establishing grooming behaviour. Rather, grooming would be determined by the record of interaction between the groomer and the target child. A distinctive feature of text content over images is that the mood, attitude and intentions of an offender may be apparent from the content of text-based interaction in a fashion that could not be matched by mere possession of extreme imagery.

One further point confirms the need to step beyond a SAP or COPINE classification for text content. Put simply, while the incidence of grooming may be disconcertingly high, the incidence of paedophiles providing a factual textual record of their child exploitation activities will be very low. Thereby, a SAP or COPINE category will rarely be usefully applicable in cases of text-based materials. Possession of textual accounts that meet such categories will not itself constitute proof of deed. Neither will such classifications assist in the identification of grooming activity.

There is concern in many quarters about the incidence of adult grooming of vulnerable minors through communication channels such as social media, chat rooms and email. Indeed, many law enforcement agencies are devoting energy to proactive monitoring of chat forums used by young people, in order to detect, track and prevent child sexual exploitation. In this section, we have argued that a content classification model such as SAP or COPINE is not suited to text-based materials

III. Reducing Observer Impact

The clear overlap between our objective and the use of image classifications (such as COPINE and SAP), is the presumption that exposure to explicit materials may be harmful to an individual. While this may not be the case for every individual who has such exposure, many will be distressed or (at least) embarrassed when confronted by sexually explicit content. This is all the more likely in a professional context in which the individual is aware that others (including peers and superiors) are aware of the situation. For want of a better term, we have called this effect 'observer impact', with the idea that the content to which the individual is exposed may have more or less emotive force and associated discomfort, distress or damage to the observer.

As noted above, in some contexts image classification schemes can reduce the observer impact by eliminating the need to observe the raw content. Circumstances may permit the novice (or the juror) to determine the seriousness of an accused's image collection through an account of the quantity and the scale values of the images therein.

In principle, we could employ such a classification scheme for textual content and on this basis reflect the 'seriousness' of an author's text output by reference to the quantity of items and their rating on this scheme, but since few instances of text generation (or text possession) are likely to be legal offences in and of themselves, we choose to focus on the issue of observer impact for such content.

Unlike the image context, with text-based materials we have the possibility of modifying the content for presentation to individuals in a manner that permits us to convey the seriousness of the content, without full exposure to the raw materials. Furthermore, in the text setting we can potentially isolate individual words, as well as multi-word units and larger discourse components. This affords a wider degree of granular analysis and potential alteration than can readily be achieved programmatically with static or moving images.

In our development of strategies for neutralising sexually explicit language, we have considered means of reducing the emotive force of sexually explicit textual content as a basis for limiting observer impact in professional, educational or training contexts. Essentially, our approach aims to vary the content of such texts so as to accommodate the desired degree of neutralisation. In section 5, we elaborate on these strategies and how they may be combined to afford a variety of end-user contexts. Firstly, we will review some relevant background work as a basis for such textual alteration.

IV. Analysis of text content

A range of automated and semi-automated techniques have been directed toward the classification of texts. In the following, we outline several approaches that are applicable to the general problem of characterising text content.

A. Text Categorization for Monitoring of Chat Rooms

Text categorization (TC) is the problem of assigning predefined categories to text documents. An important number of statistical learning methods have been applied to this problem in recent years, including regression models [6], Naïve Bayesian probabilistic classifiers [7, 8, 9], nearest neighbor classifiers [10, 11], on-line learning approaches [12, 13], decision trees [8, 9], and inductive rule learning algorithms [14, 12, 9], neural networks [15, 16]. With more and more methods available, cross method evaluation becomes increasingly important to identify the state-of-the-art in text categorization.

B. Naïve Bayes Method

The Naïve Bayes probabilistic classifier [17] is often employed as a basis for machine learning and is widely used for text categorization because of its computational efficiency and simplicity. Naïve Bayes models have been remarkably successful in information retrieval and machine learning. In the yearly TREC (The Text REtrieval Conference) evaluations, an important number of variations of Naïve Bayes models have been used, achieving some excellent results. For text categorization, some recent comparisons of learning methods have been somewhat less favorable to Naïve Bayes methods, while still showing them to produce respectable effectiveness. This may be because the larger amount of training data available in text categorization data sets favours algorithms which produce more complex classifiers, or may be because the more elaborate representation and estimation tricks developed for retrieval and routing with Naïve Bayes have not been applied to categorization.

The main idea in Naïve Bayes methods is to use relative frequencies of words in a document as word probabilities and to use these probabilities to assign a category to the document. This assumption makes the NB classifiers far more efficient than the exponential complexity of non-naive Bayes approaches because it does not use word combinations as predictors. During the learning phase, the classifier learns a set of probabilities from the training data. It then uses these probabilities and the Bayes theorem to classify other new documents.

The category of a new document is determined in two steps: (i) an estimate of the probability of the new document belonging to each class given its vector representation is calculated; (ii) the class with the highest probability is chosen as a predicted categorization.

An increasing number of evaluations of Naïve Bayes approaches on the Reuters news corpus have been published. The method is considered naive due to its assumption of word independence, i.e., the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. Apart from the fact that this assumption is false in general, the Naïve Bayes method is surprisingly effective for this specific problem. One of the confusing aspects of the recent analysis of NB methods is a non-conventional "accuracy" measure the proportion of the correct category assignments among the total of n assignments (n is the number of test documents) where each document is assigned to one and only one category.

This "accuracy" is indeed equivalent to the standard recall assuming that each document has only one correct category, and also equivalent to the standard precision under the 'one category per document' assumption on classifiers. However, it is not equivalent to the standard definition for accuracy in text categorization literature, which is the proportion of correct assignments among the binary decisions over all category/document pairs [18]. For documents with multiple categories the standard accuracy measure is well-defined while the narrowly "defined accuracy" is not. In text categorization evaluations, the latter leads to non-comparable performance measures, contributing to the difficulty of cross-collection and/or cross-method comparisons.

As has often been observed, the independence assumptions on which Naïve Bayes classifiers are based almost never hold for natural data sets, and most certainly not for textual data. This particular contradiction has motivated three kinds of research in both machine learning and information retrieval : attempts to produce better classifiers by reducing the independence assumption, modifications of feature sets to make the independence assumption more true, attempts to explain why the independence assumption is not really necessary [19].

C. K-Nearest Neighbors Method (KNN)

The K-nearest neighbors algorithm is one of the most basic learning methods and has shown to be very effective for a variety of problem domains in which underlying densities are not known. This classification method works well in the data sets with multi-modality. It has been applied to text categorization since the early days of research [11], and has been shown to produce better results when compared against other machine learning algorithms such as RIPPER [20] and C4.5 [21]. In the K-nearest neighbors approaches all documents correspond to points in an n dimensional space where n is the dimension of the document vector representation. The nearest neighbors of a document are defined in terms of the standard Euclidean distance. This method does not have a training phase and the basic computation occurs when we need to classify a new document. Based on the classification of its K nearest neighbors in the training data the classifier then classifies the new document using the Euclidean distance as a distance metric. The main weakness of this algorithm is that it uses all the features while computing the similarity between a test document and training documents. In many text data sets, relatively small number of features (or words) may be useful in categorizing documents, and using all the features may affect performance. A possible approach to overcome this problem is to learn weights for different features (or words).

D. *Variable Kernel Simulation Metric (VSM)*

VSM [22] learns the feature weights using non-linear conjugate gradient optimization. This model has a very structured approach to find weights, but requires optimization functions to be differentiable and does not have the convergence guarantees like the linear conjugate gradient optimization. RELIEF-F [23] is another weight adjustment technique that learns weights based on the nearest neighbors in each class.

Yang et al [24] used a common data set to compare the effectiveness of KNN and Naïve Bayes algorithms for classifying Web pages. They studied the usefulness of hyperlinks, metadata in classification, and content of linked documents, and found that metadata can increase the accuracy of classification by a large factor.

E. *Support Vector Machine Method (SVM)*

The support vector machine (SVM) method has been introduced in text categorization by Joachims [25], and subsequently used by Drucker et al [26], Klinkenberg and Joachims [27]. In geometrical terms, it may be seen as the attempt to find, among all the surfaces σ1, σ2, . . . in |T |-dimensional space that separate the positive from the negative training examples (decision surfaces), the σi that separates the positives from the negatives by the widest possible margin, i.e. such that the minimal distance between the hyperplane and a training example is maximum; results in computational learning theory indicate that this tends to minimize the generalization error, i.e. the error of the probability of misclassifying an unseen and randomly selected test instance. SVM methods were usually constructed for binary classification problems [28], and only recently have they been adapted to multiclass classification. One of the advantages that SVMs have to offer for text categorization is that dimensionality reduction is usually not needed, as SVMs tend to be fairly robust to over fitting and can scale up to considerable dimensionalities. Recent extensive experiments by Brank et al [29] also indicate that feature selection tends to be detrimental to the performance of SVMs. The major drawback of this method is the use of optimization techniques on text categorization which are very challenging and computationally expensive when the data set is large. There are currently several freely available packages for SVM methods.

F. *Neural network techniques (NNet)*

Neural network approaches to text categorization were evaluated on the Reuters-21450 corpus by [15] and [16], respectively. While [16] only used perceptrons, [15] tried both three-layered neural networks (with a hidden layer), and a perceptron approach (without a hidden layer). According to [30] both systems use a separate neural network per category, learning a non-linear mapping from input words (or more complex features such as singular vectors of a document space) to a category. Wiener's studies [15] suggested some advantage for combining a multiple-class NNet (for higher-level categories) and many two-class networks (for lowest-level categories), but they did not compare the performance of using a multiple-class NNet alone to using a two-class NNet for each category.

G. *Bag-of-Words*

N-gram-based Bag-of-Words features are simple yet effective means of identifying similarities between two utterances, and have been used widely in previous research on dialogue act classification for online chat conversations [31, 32]. However, chats containing large amounts of noise such as emoticons and typos pose a greater challenge for simple BoW methods. In contrast, keyword-based features [33] have achieved high performance, although such systems are more domain-dependent.

H. *Decision tree (DTree)*

DTree is a well-known machine learning approach to automatic induction of classification trees based on training data [21, 17]. Applied to text categorization, DTree algorithms are used to select informative words based on an information gain criterion, and predict categories of each document according to the occurrence of word combinations in the document. Evaluation results of DTree algorithms on the Reuters text categorization collection were reported by [8] (using the IND package) and [9] (using C4.5), respectively.

I. *Web filtering*

Web Filtering is the process of screening of Web requests and evaluation of the contents of the received Web pages to block undesired Web pages. Web filtering has the following application: (i) as the Internet is becoming an important source of information, it can host pornographic, violent and other contents that are inappropriate for most viewers, so the basic idea behind Web filtering is to protect against inappropriate and dangerous content; (ii) web filtering can be used to block access to pages that are against a defined policy. Previous Web filtering approaches include the following:

1) Blacklists and Whitelists: Blacklists and whitelists are lists of Web sites that must be allowed or blocked, respectively. Blacklists are usually created by examining and evaluating Web sites manually and deciding whether a site can be classified as a member of a forbidden class with dangerous content, such as "Nudity" and "Violence". Web sites can also be automatically included in blacklists if their domain name contains keywords like "sex" or "porn" or "xxx". In whitelisting, a list of innocent and permissible sites is generated and anything else is blocked. The main problem with both these lists is that because new Web sites continually emerge, it is hard to construct and maintain complete and up to date lists.

2) Keyword Blocking: In this approach a list of predefined keywords is used to identify undesirable Web pages. If a page contains a certain number of forbidden keywords, it is considered undesirable and dangerous. The main problem with this approach is that the meanings of words can depend on the context. As an example, Web pages about breast cancer research could be blocked because of the occurrence of the word "breast" that is used as a keyword for the "pornography class". Another problem is that the system is easily defeated using words intentionally or unintentionally misspelled. For example, a malicious site can replace the word "pornographic" with "pornogaphic" to thwart and confuse this kind of filtering system. Such replacement will have little effect on the readability of the page by human users but would make it significantly more difficult for filtering systems to correctly find the original keyword. There are two main methods of implementing Web filters: (i) implementing as a separate filter on each end-computer; and (ii) implementing as part of a firewall that controls the traffic of the network.

Lee et al [34] applied artificial neural network to filter pornographic pages. They used a collection of pornographic and non-pornographic Web sites to train artificial neural network which could be used to decide whether a given Web page is pornographic. The method requires high computation and therefore is unsuitable for real-time application.

J. Discussion

Currently, text categorization presents great challenges due to the large number of attributes present in the dataset, attribute dependency, large number of training samples, and multi-modality of categories. Existing classification algorithms address these unique challenges to varying degrees. While such approaches promise varying degrees of success in classifying texts, they are less well suited to the task of working within texts. They may have a role in deciding whether any specific text needs to be altered for the purpose of reducing observer impact, but this meta-problem will not arise in the domain under consideration, since we suppose a context in which the textual content is being deliberately presented to the subject because of its known explicit content. Chan et. al, [35] note the difficulties associated with blocking text-based pornography and comment:

> A number of approaches to identifying pornographic or obscene web pages using text alone have also been attempted, with generally poor results. Examples include: a search for strings such as "sex" – which fails to distinguish sex education or zoology from pornography; a search for obscenities, as whole or part words – fails by treating "Scunthorpe then added a fourth ..." as an obscene, rather than a soccer reference. (op. cit., p. 7).

Polpinij et. al. [36], employ a text classification based on a supervised machine learning algorithm that employs Support Vector Machines (SVMs). Their approach combines 'bag of words' with term weighting and Bayesian probability calculation of n-gram sequences (multi-word units). A variety of keywords are predefined and fed into the calculation:

> In general, pornographic phrases of bigrams often start with these words such as "adults", "sex", "free" etc. Finally, we used these term word as the features. (op. cit., p.502)

Having combined complex image filtering facilities with these text analysis techniques, the authors point out that

> it is a case of binary text classification, since it involves the classification of incoming documents into two disjoint categories: the pornographic web sites and the non-pornographic web sites. (op. cit., p.501)

Ho and Watters [37] adopt a Bayesian approach to differentiate pornographic from non-pornographic web pages, based upon the text content of the pages. While this has relevance to our concern with sexually explicit textual content, this study specifically focused on (i) 'entry pages' to sites containing adult content (specifically, pornographic images); (ii) web pages containing pornographic images, and (iii) directories of pornographic web sites (op. cit., p.4793). We can reasonably assume that the characteristics of text found in such web pages will differ from the sexually explicit textual content of text that does not accompany (or describe) associated sexual imagery.

While approaches such as [36] and [37] may assist in distinguishing two classes of web site, this is far from the level of discrimination required to isolate specific phrases or multi-word units in terms of their 'emotive strength'. In order to develop a mechanism that facilitates progressive neutralisation of sexually explicit content, we need two decision making components. The first of these is the procedure that identifies and characterises the text content (we call this 'characterisation'). The second required component is a procedure that applies the appropriate reduction in the emotive force of the sexual content (we call this 'neutralisation'). These two steps are sequential, with the neutralisation relying heavily (but not exclusively) upon the prior details of characterisation. In the following, we outline the nature of these two components, how they are interrelated and propose a range of configuration alternatives to meet differing neutralisation requirements.

K. Textual characterisation

Our requirement for text characterisation differs from the usual approaches to text classification and text categorisation in terms of the granularity. Our starting point will be sets of texts that are already classified as having sexually explicit content. Within these texts, we seek means to identify and characterise the syntactic components that contribute to this explicitness. These syntactic units (individual words and expressions) are the target for our neutralisation. This identification task may be approached in different ways that vary in their extent of automatic detection of the target content. Once the complete text has been processed and all of the components of 'sexual content' identified and annotated, the original text and its annotations serve as input to the neutralisation stage.

For the present, our work focusses on the feasibility of such neutralisation and centres on strategies for effecting such change in sexually explicit texts. For this reason, we have prototyped our approach by means of a simplified characterisation stage in which a predefined dictionary of words and phrases is pre-established as input to the identification and annotation process.

1) Characterisation data

Our starting point for characterising text content as sexually explicit was to create a dataset of sexually explicit terms. These were gathered through a survey of on-line materials, including chat room logs and sex stories. The initial source for this content was a set of chats logs from the 'perverted justice' web site (http://www.perverted-justice.com/). This site records the activities of volunteers seeking to identify and aid the prosecution of paedophiles who use online chat rooms as opportunities to groom minors for sex. One output of this work is a record of chat interactions between suspects and the volunteers (posing as minors). The language in such contexts is often highly sexualised and explicit. From this initial survey, we derived the sexually explicit vocabulary list shown in Table II.

Our vocabulary list[1] was extended by exploring other Internet sources of sexually explicit materials. A rich source of such content comes from archives of sex stories. For the most part, these texts are intended to be text-based pornography. Adding the sexually explicit language from these sources allowed us to increase our data set to include the terminology listed in Table III.

TABLE II. INITIAL VOCABULARY LIST DERIVED FROM CHAT LOGS

Nouns	Verbs	Adjectives	Phrases
Ass	Bang	Hard	Anal intercourse
Blowjob	Cum	Naked	Booty sex
Boner	Jerk off	Sexy	Cyber sex
Boobs	Masturbate	Sexier	Dip your stick
Clit	Rub		Dip your wick
Climax	Touch		Giving head
Cum	Finger		Hard cock
Cock	Fuck		Jacking off
Dick	Screw		Oral sex
Dildo	Shag		Wet dreams
Erection	Suck		
Ejaculation	Wank		
Finger			
G-spot			
Incest			
Knob			
Nipples			
Orgasm			
Penis			
Pussy			
Panties			
Rape			
Scrotum			
Sperm			
Thong			
Vagina			

TABLE III. VOCABULARY SET FROM SEX STORIES

Ass	Circle Jerk	Fagfucker	Knockers
Asscracker	Cluster fuck	Felch	Muff diving
Asshopper	Cock	Fistfucking	Nuts
Assjacker	Cocksucking	Fuckbutter	Nut sack
Ass smacking	Cornholing	Fuckstick	Prick
Assfucking	Cum	Gang Bang	Pussy
Arse	Cumshot	Gayfuck	Pussy licking
Arsehole	Cunt	Get laid	Screw
Balls	Cunthole	Gooch	Shag
Bang	Cunt lapping	Hand fuck	Slut
Blowjob	Dick	Hand job	Splooge
Butt hole	Dickmilk	Hoe	Squirting
Bollocks	Dicksucking	Jacking off	Titfuck
Bonk	Dip your stick	Jerking off	Twat
Boner	Dip your wick	Jill off	Qwif
Boobs, tits	Eating pussy	Jizz	Wang
Bumblefuck	Faggot	Knob	Wank

L. Textual neutralisation

The identification of such sexually explicit syntactic forms populates our 'first level' data set and allows us to explore the possibility of reducing the emotive force of the content. Since we anticipate contexts where the observer may be expected (or required) to grasp the underlying meaning, this should be accommodated in the degree of neutralisation. A step toward this is to devise a second level set that is considered milder in emotive force than the first level items, but does not overly conceal the meaning of the raw content. To this end, we employ an asterisk replacement

[1] The relatively small extent of vocabulary in our data set is not an issue so long as it serves in exploring the effectiveness of our neutralization strategies.

strategy to provide us with level two syntactic forms. In this case, vowels in the original syntactic form are replaced by asterisks. If the original terminology contains more than one vowel, the degree of asterisk replacement can also be varied. By default, level two replacements only substitute the first vowel in each word component in any target unit of text. For example, ‘prick’ becomes ‘pr*ck’ and ‘tits’ becomes ‘t*ts’. Our assumption is that this strategy allows us to reduce the emotive force whilst retaining the meaning.

A further step in neutralisation is achieved by establishing a level three set of replacement terms. Relative to level one terminology, level three is comprised of terms that are either ‘neutral’ or euphemistic relative to their first level associates. For instance, ‘muff diving’ appears in the first level, while ‘oral sex’ appears in the third level. Similarly, ‘fuckstick’ in level one corresponds to ‘penis’ in level three. This strategy affords a further means of varying the ‘intensity’ or explicitness of the sexual content, whilst endeavouring to retain the meaning.

Level four provides the greatest level of neutralisation is achieved by replacing every character in a level one occurrence with asterisks. This measure goes furthest in reducing the impact of the original content but risks sacrificing the meaning. However, given the context in which such replacement would take place, the observer should be in no doubt about the kind of underlying meaning that is being concealed. In level four replacements, each letter of the explicit syntax is replaced by an asterisk and any word spaces are retained. For example, the raw term ‘hand job’ would be replaced by ‘**** ***’.

Based upon our collected dataset of sexually explicit syntactical forms and these three levels of replacement, we have a simple means to identify forms of level one and replace them with ‘lesser force’ items from level two, three or four. Further examples of level one terminology and associated level two, three and four replacement are shown in Table IV.

1) Added dimensions

In order to accommodate the richness of possible sexual content as well as the variety of contexts and individuals for whom some form of neutralisation may be appropriate, the substitution based on the presumed levels of emotive force can be made more sophisticated and flexible. In the first place, our examples have so far assumed that all sexually explicit terminology (in level one) has the same initial strength (emotive force). Instead, we may attach greater weight to some content of level one above others. For example, ‘gooch’ may be regarded as stronger than ‘pussy’ although both appear in level one. This observation allows us to add a second dimension (strength) to the level one data items. Presently, we are working with three strength levels. Thereby, ‘gooch’ has a value of 3 (highest), ‘pussy’ has a value of 2 (medium) and ‘vagina’ has a value of 1 (lowest). With use of this additional dimension we can select from a wide variety of neutralisation options for any individual observer. Firstly, we can select how any level one item is replaced according to its strength (not simply by its level one status). Additionally, this permits us to ‘mix and match’ replacement strategies to accommodate different individual (or different varieties of raw materials). In this fashion, the various ‘dimensions’ derived from enriching the characterisation stage can be combined to afford a flexible set of neutralisation options. The objective is to establish a mechanism that is sufficiently flexible to meet a range of individual and situational requirements (in terms of language neutralisation).

TABLE IV. DIFFERING LEVELS OF REPLACEMENT

Level 1 (Original)	**Level 2**	**Level 3**	**Level 4**
pussy licking	p*ssy l*cking	oral sex	***** *******
hard-on	h*rd-*n	erection	****_**
hand fuck	h*nd f*ck	masturbation	**** ****

A further refinement is being considered. Each document may be gauged for the density of its sexually explicit content. There are two methods that may apply for this end. Firstly, the density may be calculated in terms of the ratio of sex words to total words[2]. Documents with a higher ratio value are more sexualised in their language. This value may be used to influence the selected replacement strategy. Secondly, we may employ an external reference source (essentially, a much larger dataset) that provides a gauge of how common is the occurrence of each sexually explicit syntactic item, relative to the external reference materials. For any considered document, we may then calculate the sum of averages across all of the sexual content items, thereby deriving a relative value that takes some account of real world usage. This may have the advantage of accounting for the likelihood that an individual will be acquainted with any specific terminology and may provide a more compelling measure for comparing documents with sexually explicit materials

Such a value, which we term the Average Sexuality Index (ASI), may be derived as shown in (1) and used to rank documents in terms of their sexual content as a precursor to selection of a corresponding strategy for neutralisation of that content. Furthermore, we may find that such sexuality index values assist in differentiating highly sexualised grooming content from other innocent text-based interactions.

$$\text{ASI} = \left(\frac{1}{n_c}\right)\left(\sum_{i=1}^{m} f_i + n_i\right) \tag{1}$$

Where:

ASI = average sexuality index

n_c *= total sexual item occurrences in sample text*

m = number of different sexual items in sample text

f_i *= frequency of sexual item i in reference corpus*

n_i *= number of instances of sexual item i in sample text*

[2] This calculation may also take account of the individual strength of each sexual content item.

M. Conclusions

This paper details the development of a mechanism for controlled neutralisation of sexually explicit language in text-based documents. The objective is to explore the application of such an approach for contexts where we wish to affect the content to which a student, trainee or other professional may be exposed. In turn, this is motivated by a desire to control the quantity and degree of such content in the hope of minimising any detrimental effects (observer impact) that such content may have on ill-prepared individuals. We have outlined our approach and the associated neutralisation strategies in terms of a selected dataset of sexual terms and phrases. We are not yet in a position to consider subtle aspects such as author's intention or dialogue constructs and this may prevent the application of our approach to richer language contexts such as online grooming, our development serves as a test-bed to explore the more limited application of the strategies outlined here. In due course, we will extend the textual identification stage to accommodate more sophisticated item matching, including syntactic variations of these terms, different tenses, spellings, hyphenations, etc., while higher level constructs such as discourse features are a longer term target for this work.

References

[1] Quayle, E. "The COPINE Project". Irish Probation Journal (Probation Board for Northern Ireland) 5, September, 2008.

[2] Taylor, M.; Quayle, E., and Holland, G. "Child Pornography, the Internet and Offending". The Canadian Journal of Policy Research (ISUMA) 2 (2): 94-100, 2001.

[3] Regina v Oliver. Court of Appeal, 2002. http://www.inquisition21.com/pca_1978/reference/oliver2002.html

[4] CEOP. Threat Assessment of Child Sexual Exploitation and Abuse, 2012. Available online at: http://ceop.police.uk/Documents/ceopdocs/CEOPThreatA_2012_190612_web.pdf [Last accessed: 19th September, 2012]

[5] Vella, P. Understanding Computer Evidence, Evidence Matters, 2012. http://www.1gis.co.uk/img/evidence.pdf [Last accessed: 19th September, 2012]

[6] Yang, Y., Chute, C.G. An example-based mapping method for text classification and retrieval. ACM Transactions on Information Systems (TOIS) 12(3):252-77, 1994.

[7] Tzeras, K., Hartmann,S. Automatic Indexing Based on Bayesian Inference Networks. SIGIR 1993: 22-34.

[8] Lewis, D., D., and Ringuette, M. 'A comparison of two learning algorithms for text categorization', in Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, (Las Vegas, US), pp. 81–93, 1994.

[9] Moulinier, I., 'Feature Selection: A Useful Preprocessing Step'. BCS-IRSG Annual Colloquium on IR Research, 1997.

[10] Creecy, R. H,. Masand, B. M, Smith, S. J., and Waltz, D. L., 'Trading mips and memory for knowledge engineering'. Communications of the ACM, 35:48-64, 1992.

[11] Yang, Y. (1994). Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In SIGIR-94.

[12] Cohen, W., W., and Singer, Y., Context-sensitive learning methods for text categorization. In SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 307-315., 1996.

[13] Lewis, D. D., Schapire, R. E., Callan, J. P., and Papka, R. Training algorithms for linear text classifiers. In Proc. 19th Annual Intl. ACM SIGIR Conf. on R&D in Information Retrieval, pages 298–306, 1996.

[14] Apté, C., Damerau, F., Weiss, S. M., Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3), 233-251, 1994.

[15] Wiener, E., Pedersen, J., O., and Weigend, A., S., A neural network approach to topic spotting. In: Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), 1995.

[16] Ng, H., T., Goh, W., B., and Low, K., L., Feature selection, perceptron learning, and a usability case study for text categorization. In: 20th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), pp. 67–73, 1997. Available online at: http://dl.acm.org/citation.cfm?id=258537. [Last accessed: 19th September, 2012]

[17] Mitchell, T., Machine Learning. New York: McGraw Hill, 1997.

[18] Harman, D., K., The First Text REtrieval Conference (TREC-1), Gaithersburg, MD 20899. National Institute of Standards and Technology. Special Publication 500-207, 1993.

[19] Lewis, D., D., Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. Vol. 1398/1998, 4.15, 1998. Available online at: http://www.springerlink.com/content/wu3g458834583125/ [Last accessed: 19th September, 2012]

[20] Cohen, W., W., Fast effective rule induction. In Proc. of the Twelfth International Conference on Machine Learning. 1995. Available online at: http://www.cs.utsa.edu/~bylander/cs6243/cohen95ripper.pdf [Last accessed: 19th September, 2012]

[21] Quinlan, J., R., C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA., 1993.

[22] Lowe, D., G., Similarity metric learning for a variable-kernel classifer. Neural Computation, vol. 7, no. 1, pp. 72-85, 1995. Available online at: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1995.7.1.72 [Last accessed: 19th September, 2012]

[23] Kononenko, I., Estimating attributes: Analysis and extensions of relief. In ECML-94: Proceedings of the European Conference on Machine Learning, Secaucus, NJ, USA, 171–182. New York: Springer, 1994.

[24] Yang, Y., Slattery, S., and Ghani, R., A study of approaches to hypertext categorization. Journal of Intelligent Information Systems, 18(2-3), 219-241, 2002.

[25] Joachims, T., Text categorization with Support Vector Machines: Learning with many relevant features, in Proceedings of ECML-98, 10th European Conference on Machine Learning (C. N´edellec and C. Rouveirol, eds.), no. 1398, (Chemnitz, DE), pp. 137–142, Springer Verlag, Heidelberg, 1998.

[26] Drucker, HD., Wu, D., and Vapnik, V., Support Vector Machines for Spam Categorization. IEEE Transactions On Neural Networks,10(5):1048-1054, 1999.

[27] Klinkenberg , R., Joachims, T., Detecting Concept Drift with Support Vector Machines, Proceedings of the Seventeenth International Conference on Machine Learning, p.487-494, June 29-July 02, 2000.

[28] Vapnic, V., The Nature of Statistical Learning Theory. Springer, New York, 1995.

[29] Brank, J., Grobelnik, M., Milic-Frayling, N., Mladenic, D. Feature selection using support vector machines. In: Proc. Third Internat. Conf. on Data Mining Methods and Databases for Eng. Finance Other Fields, pp. 25–27, 2002.

[30] Yang, Y., An evaluation of statistical approaches to text categorization. In: *Information Retrieval Journal*, 1 (1-2), 1999, pp. 69–90.

[31] Louwerse, M.M., Crossley, S.A. Dialog act classification using n-gram algorithms. In Proceedings of the 19th International Florida Artificial Intelligence Research Society, 2006.

[32] Ivanovic, E., Automatic instant messaging dialogue using statistical models and dialogue acts. Master's Thesis. The University of Melbourne, 2008. Available online at: http://repository.unimelb.edu.au/10187/2820 [Last accessed: 19th September, 2012]

[33] Forsyth, E., N., Improving Automated Lexical and Discourse Analysis of Online Chat Dialog. Master's thesis. Naval Postgraduate School, 2007. Available online at: http://www.ldc.upenn.edu/Catalog/docs/LDC2010T05/Forsyth_thesis.pdf [Last accessed: 19th September, 2012]

[34] Lee, P., Y., Hui, S., C., and Fong, M., Neural Networks for Web Content Filtering. Intelligent Systems, 17(5):48-57, 2002.

[35] Chan, Y., Harvey, R., Smith, D.: Building systems to block pornography. In Eakins, J., Harper, D., eds.: Challenge of Image Retrieval, BCS Electronic Workshops in Computing series, 34–40, 1999.

[36] Polpinij, J., Sibunruang, C,. Paungpronpitag, S., Chamchong, R., Chotthanom, A., 'A web pornography patrol system by content-based analysis: In particular text and image,' IEEE International Conference on Systems, Man and Cybernetics, 2008. 500-505, 12-15 Oct. 2008.

[37] Ho, W.H.; Watters, P.A.; , Statistical and structural approaches to filtering Internet pornography, *IEEE International Conference on Systems, Man and Cybernetics, 2004*, vol.5, pp.4792-4798.