Measuring Sexually Explicit Content

Katherine Darroch and George R S Weir[1]

Department of Computer and Information Sciences
University of Strathclyde

george.weir@strath.ac.uk

Abstract: In this paper we describe an experiment to investigate methods of measuring sexually explicit content in text documents. As a starting point, sample data was collected from a variety of sources and manually sorted into three categories: (i) sexually explicit, (ii) non-sexually explicit, and (iii) content that contained sexually explicit terms but was not sexually explicit - e.g., information used for sex education. This selection of data was used as a training set in developing three software metrics of the type often used in content filtering. Thereafter, a test set of six files was used in the experiment. These test files were scored for sexually explicit content by participants in the study and by the three different metrics. The assigned scores were then compared to consider how far the metrics and the users agreed on their view of sexually explicit content. In addition to our contrast between software metrics and users, we also note interesting trends from the participant demographics.

1 Introduction

Although Internet content filtering systems are often relied upon to protect young people from exposure to sexually explicit content, these systems are not perfect (Ding, Chi, Deng, & Dong, 1999). They may over block content with the effect that perfectly safe and appropriate content, including educational material, is restricted. On the other hand, such systems may struggle to filter content correctly, leaving potentially sexually explicit websites available to all (Nicoletti, 2013). Such difficulties often stem from the inability of software systems to appreciate the context and structure and significance of textual content (Ding, Chi, Deng, & Dong, 1999).

2 Sexually Explicit Content

Sexually explicit content is defined as presentation material that contains “depictions or descriptions of…sexual references, full or partial nudity, bestiality, sexual acts... [and] sexual behaviour that has a violent context” (Leibowitz, Harbour, Kovacic, & Rosh, 2009, p. 2). Sexually explicit content can take many forms including sexually explicit language, images and videos and the content can vary in the severity of explicitness (Sirianni & Vishwanath, 2012). Defining sexually explicit content is difficult as not everyone agrees on what is sexually explicit and what is not. Furthermore, domestic laws on what is sexually explicit can also differ. On both a global and domestic scale sexually explicit content can be hard to quantify and categorise (Lloyd, 2011, pp. 258-261). However, the general consensus is that such material is adult-oriented content from which children and young people should be protected (Savirimuthu, 2012, p. 192). In order to achieve this, parental controls and content filters have been designed and implemented to prevent children from accessing this material by blocking the websites that contain sexually explicit content (Savirimuthu, 2012, pp. 242-246).

2.1 Content Filtering

Content filtering is the use of a program (a content filter) to screen and block web pages or emails that are unwanted by the user (Ding, Chi, Deng, & Dong, 1999). Content filters can be installed on servers, individual computers, implemented by Internet Service Providers (ISPs) and within a network. Software filters can be bought or downloaded and installed on a computer (Lee, Hui, & Fong, 2002). The filter acts as a middle man between the user’s computer and the internet reviewing the content that is accessed. The software filter can then prevent the content from being viewed. Hardware filters are filters located within the hardware used to connect a computer to the internet and can be effective in blocking content before it reaches the user’s computer by screening the content while it is in transit. ISPs can also provide content filters as part of their service (Nicoletti, 2013). Content filters are implemented in two main ways, using URL lists and content

[1] Corresponding author.

inspection. URL filtering works by blocking sites that are contained on a black list of URLs. When someone attempts to access a website on the black list that access is denied. Content filters that inspect the content of the site to look for keywords work by searching the site for keywords or phrases that would be linked with unwanted content and block access to sites which contain these keywords (Hidalgo, Sanz, Garcia, & Rodriguez, 2009).

The method of content filtering that I will be basing my project on will be software based content inspection filtering that can rate content based on keywords as this will allow me to evaluate the effectiveness for these systems to recognise sexually explicit text.

Content filtering systems can be used to block a variety of content including but not limited to sexually explicit content, spam and advertisements (Deibert, Palfrey, Rohozinski, Zittrain, & Stein, 2008).

2.1.1 Content Labelling

Content Labelling is a method that allows creators of websites to self-assess their web sites content by completing a questionnaire. The results of the questionnaire can then be used to label the content. Once labelled these sites can be blocked or restricted by content filtering systems that use these labels. However as the websites content is self-regulated it is possible for sites to labelled incorrectly or not labelled at all (Hidalgo, Sanz, Garcia, & Rodriguez, 2009).

2.1.2 Email Filters

Email filters are predominately used to filter SPAM messages and can be found on almost all email applications. SPAM messages can contain content that is adult in nature including sexually explicit content (Organisation for Economic Co-operation and Development, 2006). Bayesian spam filtering is a common filtering technique that works by calculating the probability that any given message is spam based on the words contained in the message (op. cit.). Therefore it can often be the case that emails that are not spam are classed as spam and vice-versa. Many other techniques as well as the Bayesian approach are used to filter SPAM messages with differing levels of success.

(i) Microsoft Smart Screen

Microsoft Smart Screen is a commonly used email filter designed for Windows and its accompanying software including Outlook and Hotmail. Microsoft Smart Screen is a Bayesian spam filter (Organisation for Economic Co-operation and Development, 2006) that has been designed to determine if emails are legitimate or spam by gathering data from Windows Live Hotmail users. Microsoft claims that in using this software 95% of all spam is blocked (Microsoft, 2014).

(ii) Advertisement filters

Advertisement filters are content filters that block advertisements that appear when browsing the internet. These adverts can contain sexually explicit content. Advertisement filters work by blocking pop-up windows that come from certain web addresses. In order to do this the filtering system maintains a URLs list containing ads that should be blocked and prevents these URLs from displaying advertisements (Singh & Potdar, 2009).

(iii) AdBlock Plus

AdBlock Plus is one of the most popular advertisement filtering/blocking systems on the internet today. AdBlock Plus is not a standalone system but instead a browser extension that can be installed on a browser to run behind the scenes. AdBlock Plus works by only blocking what it is told to block. The user tells the filter what to block by signing up to different URL lists that they wish to filter against. (Singh & Potdar, 2009).

(iv) ISP Level Content Filters

ISP content filters exist to restrict/block access to various forms of content not just emails or ads across a whole network (Hidalgo, Sanz, Garcia, & Rodriguez, 2009). Instead ISP content filters are created under instruction from the government to block in particular illegal content such as child pornography (op. cit.). However since then the use of ISP filters have increased to now block websites that facilitate copyright infringement (Stalla-Bourdillon, 2013). ISP content filters several techniques to filter content and restrict access to sites including use URL blocking to filter content and restrict access to certain sites. URL blocking works by creating a blacklist (a list of websites that should be blocked) and preventing users from accessing these sites (Varadharajan, 2010).

(v) Cleanfeed

Cleanfeed is an ISP content blocking system developed by British Telecom, designed to block illegal websites that were listed by the Internet Watch Foundation (IWF) (Nicoletti, 2013). Cleanfeed works by checking the URL's that are requested by any given internet user are on IWF list. If they are access to the site will be restricted and an error message will appear otherwise users will be able to browse their requested URL (Akdeniz, 2010). However since its development there have been several calls for BT to adapt Cleanfeed to block sites that contain or facilitate the sharing of copyrightable material illegally. Following court action in 2011 BT were ordered to begin blocking the popular file sharing index NewzBin (Hornle, 2012).

3 Data Gathering

To secure content for our experiment, two approaches were adopted. The first was to visit websites and manually select text. The second approach was to use cURL to automate the retrieval process from specified Web addresses (Stenberg, 2014). Websites were selected randomly from Google search results for sites that would contain varying levels of sexually explicit. In order to operate with consistent experimental data, the data files had to be the same length and be in a similar style and format. Therefore all the text files selected had to be not only in the same font etc. but also in a similar writing style.

3.1 Content of Text files

To evaluate the effectiveness of the example metrics the experiment considered the ability of the metrics to differentiate between sexually explicit content and non-sexually explicit content as well as recognising context for content that contains sexually explicit content but is not actually sexually explicit. For the experiment to achieve accurate results we could not simply feed the program with only sexually explicit data. Rather, we required experimental data that contained varying degrees of sexually explicit content. In order to create a fair distribution of different levels of sexually explicit content the experimental data files had to equally represent our three categories of (i) sexually explicit data, (ii) non-sexually explicit data, and (iii) data that contained sexually explicit references but in context was not lewd or salacious.

4 Metrics

To select appropriate metrics for our research, we explored related academic work on the use of different techniques to measure sexually explicit content and shortlisted metrics that had previously shown some aptitude in recognising and scoring such content. In order to ensure that the implemented metrics work correctly a proportion of the experimental data gathered was used to test the program and then act as a control for the actual experiment and allow the results to be evaluated. To this end, the experimental data used for testing was categorised as sexually explicit or non-sexually explicit before being evaluated by the metrics. The results of the metrics were then compared to the manual categorisation to determine whether or not the metrics were performing reasonably. Thereafter, the pre-categorised data could be used to train the metrics and act as a control to compare with the experimental results. With this in mind, the experimental design developed for this research is detailed below.

1. Selection of Test Data
   The test data comprised three sets of sample data, sexually explicit data, non-sexually explicit data and sexually explicit data mixed together, and non-sexually explicit data. These data samples were in standardised format with standard length and style in order to minimise random influences on the experimental results.

2. Selection of automated metrics
   A selection of three metrics was made, based upon the apparent popularity of the content rating/filtering systems often used to block content.

3. Rating text content
   The experiment was conducted in two stages. In the first stage the test data was examined manually to determine whether it contained explicit content, before highlighting the words or phrases that were

explicit and giving a manually-estimated rating. Subsequently, these test data samples were fed to the metrics in order to compare the results. This first stage of the experiment was used initially to ensure that the implemented metrics were effective as well as acting as control data for the second stage. In the second stage of the experiment, a sample group of volunteers was asked to gauge the sexually explicit language within the test data and rate this using a slider.

4. Implementation of experiment using metrics
   Each of the metrics was applied to the same test data as rated by the volunteer group. The results of applying each metric was output in rank order of explicitness. This allowed for easy comparison with the rankings assigned by the volunteers.

5. Comparison and Evaluation of Results
   The measures of explicitness assigned to the test data samples by each metric was compared to the responses from the volunteers. The alternative results comparison was based upon how close the ratings given by the approach are to the rating given by a user and by whether or not the approach achieved that rating by selecting the same terms that the user did when deciding on a rating for the test data. The approaches that are closer to the results given by the user are considered more successful at measuring sexually explicit content. In addition, the approaches were evaluated on their ability to distinguish between explicit and non-explicit text, context of the sentences that make them explicit or non-explicit and their ability to recognise spelling mistakes and minimise over blocking.

4.1 Bayes Theorem

Bayes theorem defined by 18th Century mathematician Thomas Bayes is a “theorem describing how the conditional probability of each of a set of possible causes for a given observed outcome can be computed from knowledge of the probability of each cause and the conditional probability of the outcome of each cause” (Oxford Dictionaries, 2014). The formula for Bayes theorem is shown in Figure 1 (Statistics How To, 2014), the formula denotes that the conditional probability of A given B can be calculated based on the knowledge that B has happened or is true (Cornfield, 1967, p. 35). In order for the formula to work it must be trained. Training is done by giving the algorithm training data which is manually categorised as A or B from this data Bayes is able to calculate probabilities of A given B or B given A by using the information in the training data. The more training data given to the Bayes Theorem the more effective it is in determining accurate probabilities (Process Software, 2014). Bayes theorem was chosen as metric to evaluate the effectiveness of content filtering systems as it is a very flexible yet simple metric that can be applied to almost any situation (Zellner, 2007, pp. 14-15).

$$\boxed{P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}}.$$

Figure 1: Bayes Theorem

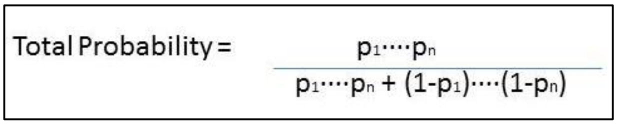
4.1.1 Bayes Theorem for Measuring Sexually Explicit Content

In order to apply Bayes theorem to assess the probability that a text file was sexually explicit, the probability that each word in the text file was sexually explicit had to be calculated and combined to give the overall probability. To do this the original Bayes formula had to be amended to apply to sexually explicit content. The amended formula that was created can be seen in Figure 2 where P(E|W) is the probability that a file is sexually

$$\boxed{P(E|W) = \frac{P(E) \cdot P(W|E)}{P(E) \cdot P(W|E) + P(W|NE) \cdot P(NE)}}$$

Figure 2: Adapted Bayes equation

explicit given the word is in it, P(E) is the overall probability that any given file is sexually explicit, P(W|E) is the probability that the word appears in sexually explicit documents, P(W|NE) is the probability that the word appears in non-sexually explicit text and P(NE) is the overall probability that any given file is not sexually explicit. This formula was derived using the Bayesian Spam filtering formula as a guide (Process Software, 2014). The formula detailed above only calculates that a given text file is explicit given that a word is contained within it therefore the probabilities calculated for each individual word in the file must be combined to reveal the overall probability that the text file is sexually explicit given all the words that are contained in it.



$$\text{Total Probability} = \frac{p_1 \cdots p_n}{p_1 \cdots p_n + (1-p_1) \cdots (1-p_n)}$$

Figure 3: Combining probabilities

The probabilities were combined using the formula in Figure 3 that is commonly used for combining probabilities in spam (MathPages, 2014).

4.2 Similarity Measure

A similarity measure indicates the similarity or distance between objects in order to divide data into clusters or categories. Generally this is done using a mathematical formula or function to calculate the similarity of objects using their descriptions or content into a value. Many similarity measures are used to compare text. Metrics for calculating similarities include Euclidean Distance and Cosine Similarity. Similarity Measures are commonly used to categorise the web and cluster similar documents or webpages in order to create efficient searches for content (Huang, 2008). The ability for similarity measures to recognise sexually explicit content and categorise correctly is an important task when categorising the web and implementing content filters. Therefore it seemed appropriate to include such a measure to determine how effective a similarity measure could be at recognising what was sexually explicit by comparing how similar a given file was to sexually explicit content. Furthermore the similarity value calculated could also be used to indicate how sexually explicit a file is based on how similar it was to other sexually explicit files.

4.2.1 Implementation of Similarity Measure for Measuring Sexually Explicit Content

Similarity Measures work by taking a collection of documents/objects that have already been categorised or clustered and comparing those documents to new ones normally using a metric or mathematical formulae (Huang, 2008). However when researching similarity metrics many of the most effective ones seemed very complex and would have required modification of the existing system to work effectively. Therefore a simpler similarity measure function was derived. The similarity measure used with the project works by measuring how similar text files are to one another by counting the number of similar words between the files. The file that is most similar to the file being scored is the file with the highest number of similar words. The similarity measure then gets the classification from the file that is the most similar and applied that to the file being scored. Although this is a very simple implementation of a similarity measure it does follow the underlying principle of the approach allowing for the effectiveness of similarity measures to measure sexually explicit content to be evaluated.

4.3 DansGuardian

DansGuardian is an open source web content filter that was developed for Linux based machines by Daniel Barron. It uses many different methods of content filtering however for this project its method for “phrase matching” to filter content was implemented within this project (Barron, 2014). The system itself is coded in C++ (Barron, How I Did It, 2014) and the source code is free to download for non-commercial use (Barron, Copyright and Licensing for DansGuardian, 2014). DansGuardians phrase matching method works by searching the document/web page for the phrases contained in the DansGuardian dictionary. Each word/phrase in the dictionary is given a score depending on how explicit the word/phrase is. By using the dictionary and totalling each phrase score a final score for how explicit the web page/document is returned allowing users to filter pages with high scores (DansGuardian, 2014) . DansGuardian was selected for this project as it is a fully functional content filtering system that is used both commercially and on personal computers. It was also selected as unlike the other metrics chosen it works by giving certain words and phrases weighted scores which in some case allows for context to be taken into consideration. For example the difference between the word sex and the phrase sex education.

4.3.1 Implementation of DansGuardian for Measuring Sexually Explicit Content

In order to implement DansGuardian into the project to evaluate its effectiveness at measuring sexually explicit content the source code had to be used as base and recoded within java to show DansGuardians basic functionality with phrase matching. In order to do this the open source code that was downloaded was examined to discover how it worked. Following this a sample of the DansGuardian Phrase Dictionary had to be extracted and included within the project. Only a sample of the sexually explicit words and phrases contained in the DansGuardian dictionary were copied into the dictionary within the project.

5 Experimental Results

This chapter will explore the results from the experiment to evaluate the effectiveness of the metrics used in content filters at recognising and measuring sexually explicit content. The Chapter will detail how results can be interpreted, what the results reveal generally and will compare the results of the metrics to each other and the users results averaged both generally and by age and gender.

Filename	*Category of file*
d1.txt	Non-Sexually Explicit
d2.txt	Contains sexually explicit language but is Non-Sexually Explicit
d3.txt	Sexually Explicit
d4.txt	Sexually Explicit
d5.txt	Non-Sexually Explicit
d6.txt	Contains sexually explicit language but is Non-Sexually Explicit

Table 1: Names and categories of the experimental files

5.1 Categories of Files used in Experiment

The files used in the experiment were manually categorised before they were used in the experiment. Table 1 reveals the filenames and categories of the experimental files. Although the categories of the files were determined before the experiment they were displayed to the participants in a random order so that there was no bias placed on the files by the project author during the study.

5.2 Interpreting the results

Each of the metrics return results for the text files in different ways depending on the calculation that they use however all the scores returned can be used to indicate whether a file is sexually explicit or not and to indicate how sexually explicit files are.

5.2.1 Bayes Metric

The Bayes metric returns the probability that the given file is explicit based on its training data. The probability returned will be a value between 0 and 1 (Peter M. Lee, 1989, p. 4). For the experiment if the probability returned is less that 0.5 then the file being scored is non-explicit, if the probability is 0.5 the file being scored is neither sexually explicit or non-sexually explicit and if the probability returned is greater than 0.5 the file being scored is sexually explicit. Although the raw probability score only indicates the category of the file it is possible to use the probability scores returned to indicate how sexually explicit a given file is. For example if a probability score of 0.8 was returned for one file and a probability score of 0.6 was returned for another file. It can be inferred that the first file has a greater level of sexually explicit content that the second.

5.2.2 Similarity Measure

The Similarity Measure returns the number of similar words that are contained within the text fie being scored and the most similar file within the training data (the similarity score). The category of this file is then decided based on the category of the most similar file. For example if the most similar file is sexually explicit then the text file being scored will be categorised as explicit also. However the similarity scores can also be used to indicate how sexually explicit a file is as the more similar words found the more sexually explicit the file is. For example a file which has a similarity score of 182 when compared to sexually explicit content will be more sexually explicit than a file which has a similarity score of 90 when compared to sexually explicit content.

5.2.3 DansGuardian Metric

The DansGuardian metric returns a score indicating how sexually explicit a file is based on the weighted scores given to each word/phrase within the DansGuardian dictionary. The higher the score returned the more sexually explicit the file is. DansGuardian phrase matching part of the filter does not directly categorise the files however from the scores returned the files can be categorised based on how high the score returned is. For example a score of 200 is likely to be sexually explicit whereas a score of 20 is not.

5.2.4 Average Score from Participants

The scores by the participants were recorded using a slider that was to be positioned by the participant to reveal how sexually explicit the given text files were. The slider was programmed to work on a scale between 0 and 100 where 0 represented a file that was non-explicit and 100 represented a file that was extremely sexually explicit. The scores given by the participants fell between those two points for all the text files. In order to compare the results these scores given by the 25 participants was averaged to give one set of results for the study. The results themselves only indicate how sexually explicit the files are however it is possible to categories the files using the score by using a similar approach to how the Bayes metric's results are evaluated. If the participants score is <=40 then the file is non-sexually explicit, if participants score is between 41-59 it is a file that contains some sexually explicit content but is no very sexually explicit and if the participants score is >=60 then it is sexually explicit.

5.3 Experimental Results

5.3.1 Bayes Metric Results

From Figure 4, we can see that the Bayes metric performed well in the experiment correctly identifying the two sexually explicit text files d3 and d4 and determining that the other files were not in fact sexually explicit. Although the Bayes metric was trained using a small data set the metric did not calculate that non-explicit files that contain sexually explicit data were likely to be sexually explicit overall. This shows that if training with appropriate data that the Bayes metric is capable of filtering sexually explicit content without over-blocking. However the result of d5 although a low probability should have been substantially lower as the file was completely non-explicit this is likely down to the limited set of data used to train the metric highlighting an important flaw/Achilles heel with the Bayes metric which is that its results depend solely on the data it has been trained with (Maeda, Yoshida, & Matsushima, 2009).

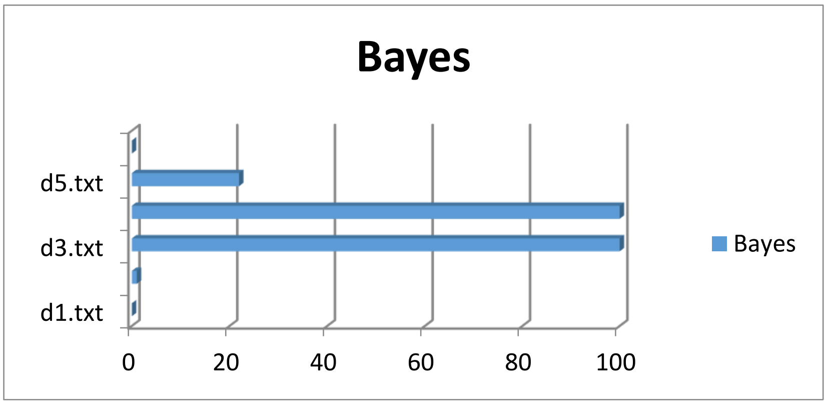


Figure 4: Results for Bayes metric

In order to compare the Bayes Metric results to the results from the other metrics the probability scores returned were multiplied by 100.

5.3.2 Similarity Measure Results

Reviewing the results of the experiment, we find that the Similarity Measure performed well in correctly identifying the two most sexually explicit files d4 and d3. However, unlike the Bayes metric, the scores given by the similarity measure show that according to its calculation d4 is more sexually explicit than d3 (Figure 5).

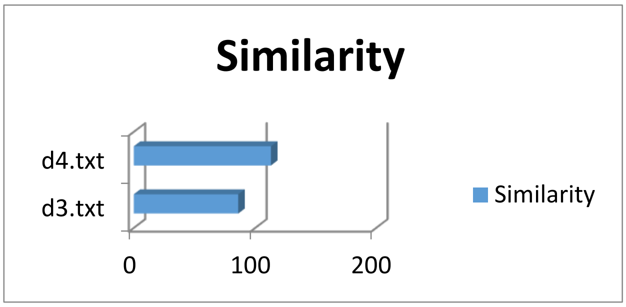


Figure 5: Similarity measure results for files d3 and d4

The similarity Measure categorised all the other files as being non-sexually explicit, returning scores for how similar to non-explicit files they were. Examining the results (Figure 6) we see that the similarity scores for the non-explicit file containing sexually explicit content (d6 and d2) are given smaller scores than non-explicit files. This is likely due to the fact that the non-explicit files will have more words in common with the non-explicit training data. This reveals that the similarity measure is able to recognise files that contain a mixture of sexually explicit content and non-explicit content and be used to filter them appropriately.

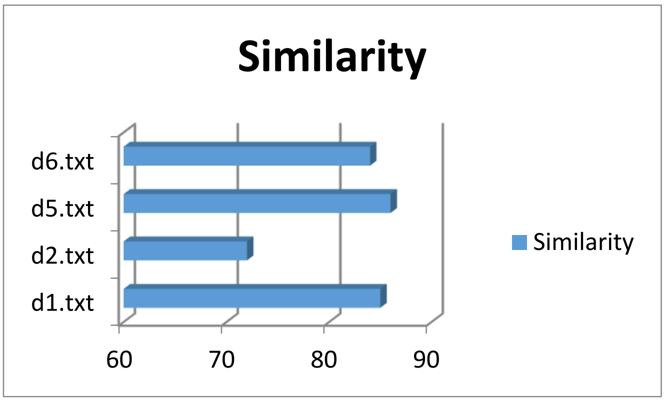


Figure 6: Similarity measure results for non-sexually explicit files

5.3.3 DansGuardian Results

Examining the experimental results for DansGuardian (Figure 7) we see that the DansGuardian-based phrase matching did not perform well. Although this metric did score the sexually explicit file higher than the non-explicit file containing sexually explicit terms it also rated the non-explicit files very highly. The reason for these results is likely down to the sample of the DansGuardian dictionary that was used in the implementation of the metric. The dictionary works by giving positive scores for sexually explicit words and negative scores for words that are non-explicit. However whether or not the score is calculated accurately depends on whether the words in the file are contained in the dictionary. As the full dictionary was not used it is possible that not all the words were given a score resulting in the DansGuardian score being inaccurate.

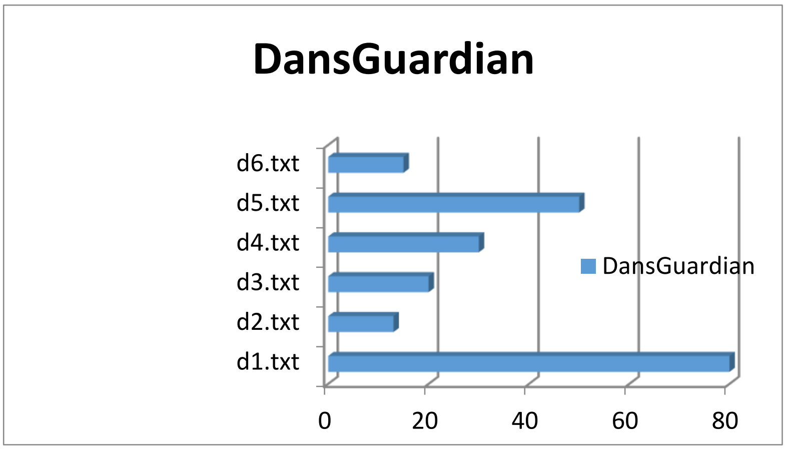


Figure 7: DansGuardian results

5.3.4 *Average Participant Results*

The participants on average scored d4 as the most sexually explicit and d3 as the next most sexually explicit file correctly identifying that they were the two explicit files (Figure 8). The participants on average were also able to identify the files that were non explicit but contained sexually explicit content this can be seen from the higher scores given to d6 and d2 than the scores give to the completely non-explicit files d1 and d5.

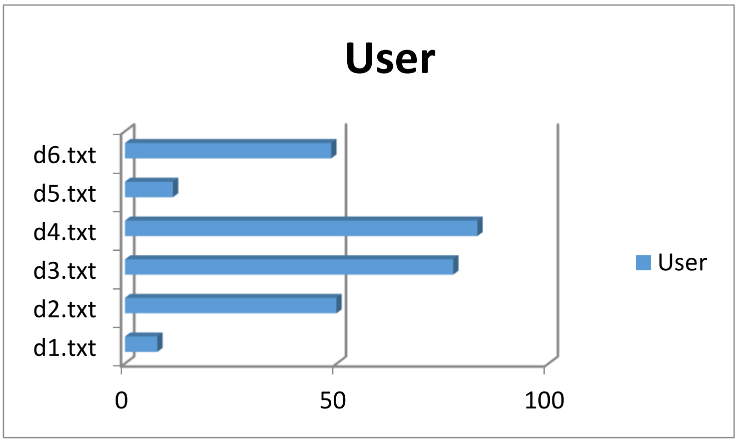


Figure 8: Participant results

The average participant scores calculated on different demographics including age and gender showed some interesting results.

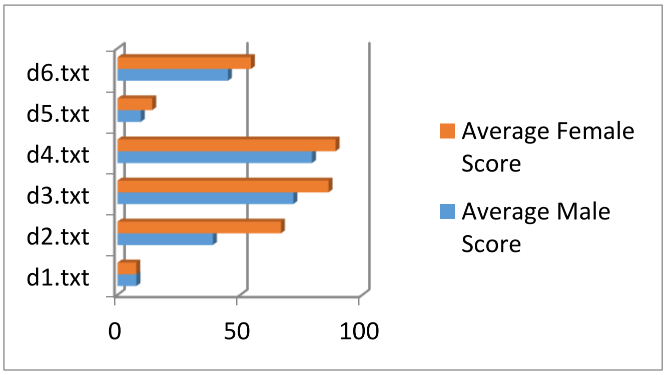


Figure 9: Results by gender

Figure 9 shows that on average female participants scored content more highly than their male counterparts.

In Figure 10 we can see that participants in the 18-25 and 36+ age group score the content much more highly than those in the 26-35 age group. These differences in participant scores reveal that the measuring of sexually explicit content is affected by several different factors related to a person's demographic.

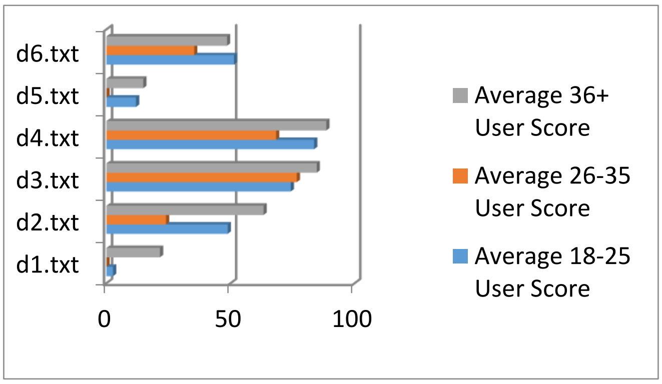


Figure 10: Results by age

5.5 Comparison of Metrics Results

From the metric results comparison displayed in Figure 11, we see that the Bayes metric and the similarity measure are both successful in identifying files d3 and d4 as the most sexually explicit. However, the similarity measure is better at indicating the level of sexually explicit content as it determines that d4 is more explicit than d3. The DansGuardian metric does this as well as it rates d4 higher than d3. However, unlike the other metrics, the DansGuardian metric does not accurately identify that they are the sexually explicit files. All of the metrics give d5 - which is pre-determined as non-explicit - higher scores than would be expected, indicating that it could be explicit. However when the results are examined carefully we can see that the Bayes score is in fact just a low probability that it is explicit rather than a negligible probability as with the other non-explicit files. In terms of DansGuardian it rated all the non-explicit files higher than would be anticipated due to the lack of non-explicit words contained in the files being found in its dictionary. The similarity measure scores for the non-explicit files cannot be compared in this way as they actually indicate how similarity the files are to non-explicit data not sexually explicit data.

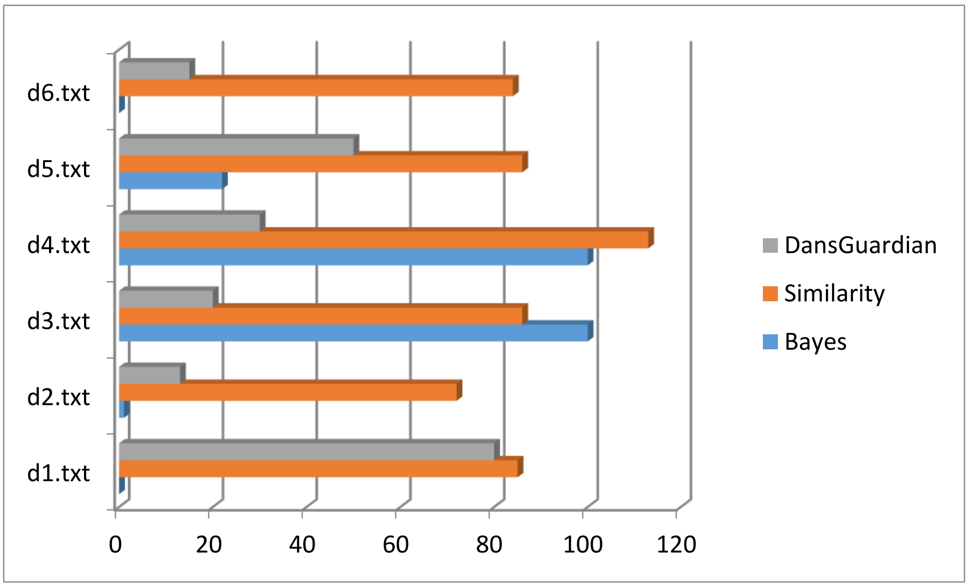


Figure 11: Metric results comparison

Following this comparison the similarity measure and Bayes metric are shown to identify the most explicit files correctly and the similarity measure and DansGuardian are better equipped to reveal the level of sexually explicit content contained in the files. In consequence, there is no clear metric that stands out as being the best at measuring sexually explicit content. Instead the effectiveness of the metrics depends on the training data used and how they are implemented in a content filtering system. However on the face of it the Similarity Measure and Bayes Metric are better at identifying sexually explicit content than DG.

5.6 Participant Results vs Metric Results

From the results displayed in Figure 12, we see that the user scores are similar to both the Similarity Measure and the Bayes scores, indicating that d4 and d3 are the most explicit files. Like the Similarity scores and the DansGuardian scores, the user scores show that d4 is more sexually explicit than d3 indicating that the Similarity score and DansGuardian are more effective than Bayes at measuring the levels of sexually explicit content within the files. However unlike the metrics which rate d5 the highest of the non-explicit files, the user results indicate that d2 - although not explicit - has more explicit content than the other non-explicit files. This may reveal that although the metrics can be used to measure the levels of sexually explicit content, they struggle with context. Although the Similarity Measure scores for the non-explicit files are not directly comparable to the scores for the non-explicit files as explained above in general the scores given by the similarity measure are the closest to the user scores. Furthermore with the exception of Bayes the user scores are significantly less than the score given by the metrics indicating that the metrics have the ability to score explicit words higher than necessary again this may be due to the metrics inability to deal with context however it could also reveal the reason behind over blocking of content due to filtering systems.

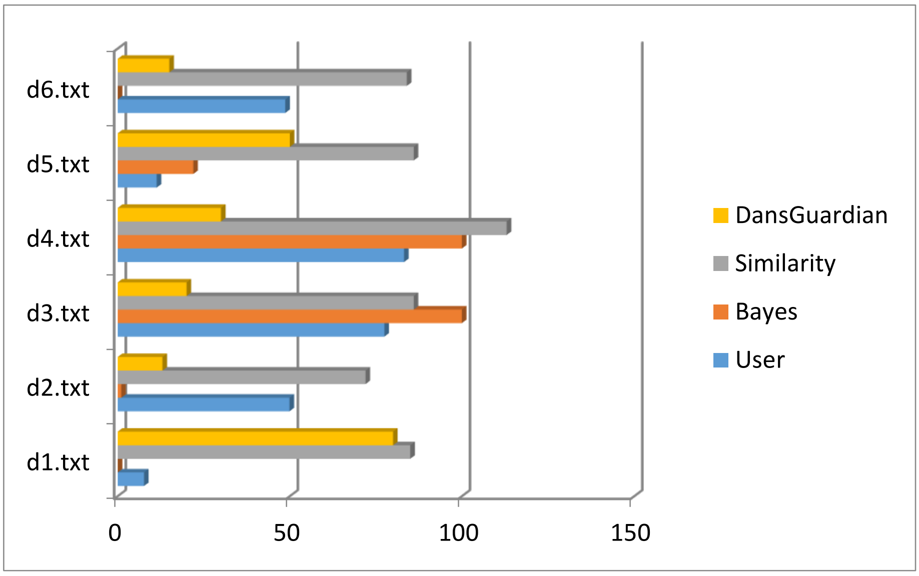


Figure 12: Comparison of metrics and users

Following this comparison the Similarity Measure and Bayes Metric identify the most explicit files correctly when compared to user results and the Similarity Measure and DG are better equipped to reveal the level of sexually explicit content contained in the files although in places they do seem to struggle with context which can be seen when the scores are directly compared to those given by the participants. Therefore there is no clear metric that stands out as being the best at measuring sexually explicit content when compared with a human being. Instead the effectiveness of the metrics depends on the training data used and how they are implemented in a content filtering system. However on the face of it the Similarity Measure and Bayes Metric are better at identifying sexually explicit content than DG.

References

Akdeniz, Y. (2010). To block or not to block: European approaches to content regulation, and implications for freedom of expression. *Computer Law & Security Review*, 260-272.

Barron, D. (2014). *Copyright and Licensing for DansGuardian*. Retrieved March 24, 2014, from DansGuardian: http://dansguardian.org/?page=copyright2

Barron, D. (2014). *How I Did It*. Retrieved March 24, 2014, from DansGuardian: http://dansguardian.org/?page=howididit

Barron, D. (2014). *What is DansGuardian*. Retrieved March 24, 2014, from DansGuardian: http://dansguardian.org/?page=whatisdg

DansGuardian. (2014). *Phraselists*. Retrieved March 24, 2014, from DansGuardian Documentation Wiki: http://contentfilter.futuragts.com/wiki/doku.php?id=phraselists

Deibert, R., Palfrey, J., Rohozinski, R., Zittrain, J., & Stein, J. (2008). Access denied : the practice and policy of global Internet filtering. In R. Deibert, J. Palfrey, R. Rohozinski, J.

Ding, C., Chi, C.-h., Deng, J., & Dong, C.-L. (1999). Centralized Content-Based Web Filtering and Blocking: How Far Can It Go? *Systems, Man and Cybernetics, 1991. IEEE SMC'99 Conference Proceedings, Vol 2*.

Hidalgo, J. M., Sanz, E. P., Garcia, F. C., & Rodriguez, M. D. (2009). Web Content Filtering. *Advances in Computers, 76*.

Hornle, J. (2012). Premature or stillborn? - The recent challenge to the Digital Economy Act. *Computer Law & Security Review, 28*(1), 83-89.

Huang, A. (2008). Similarity Measures for Text Document Clustering. *New Zealand Computer Science Research Student Conference*, (pp. 49-56). Christchurch.

Lee, P. Y., Hui, S. C., & Fong, A. C. (2002). Neural Networks for Content Filtering. *Intelligent Systems, 17*(5).

Leibowitz, J., Harbour, P. J., Kovacic, W. E., & Rosh, J. T. (2009). *Virtual Words and Kids: Mapping the Risks , A Report to Congress.* Federal Trade Commission.

Lloyd, I. J. (2011). *Information Technology Law* (6th Edition ed.). Oxford: Oxford University Press.

Microsoft. (2014). *Help keep Spam out of your inbox*. Retrieved March 22, 2014, from Mircosoft Safety and Security Centre: http://www.microsoft.com/en-gb/security/online-privacy/spam-prevent.aspx

Nicoletti, P. (2013). Chapter e66 : Content Filtering (Online Resource). In J. Vacca, *Computer and Information Security Handbook* (2nd Edition ed.). Waltham: Elsevier.

Organisation for Economic Co-operation and Development. (2006). *OECD Anti-Spam Toolkit of Recommended Policies and Measures.* Paris: OECD Publications.

Oxford Dictionaries, O. (2014). *Definition of Bayes' Theorem*. Retrieved March 23, 2014, from Oxford Dictionaries: http://www.oxforddictionaries.com/definition/english/Bayes'-theorem

Process Software. (2014). *Process Software.* Retrieved March 24, 2014, from Introduction to Bayesian Filtering: http://www.process.com/psc/fileadmin/user_upload/whitepapers/pmas/intro_bayesian_filtering.pdf

Savirimuthu, J. (2012). *Online Child Safety Law, Technology and Governance.* Basingstoke: Palgrave Macmillan.

Singh, A. K., & Potdar, V. (2009). Blocking Online Advertising - A State of the Art. *Industrial Technology, 2009 1CIT 2009. IEEE International Conference*.

Sirianni, J. M., & Vishwanath, A. (2012). Sexually Explicit User-Generate Contet: Understanding Motivations and Behaviours using Social Cognitive Theory. *Journal of Psychosocial Research on Cyberspace, 6(1) Article 7*.

Stalla-Bourdillon, S. (2013). Online monitoring, filtering,blocking......What is the difference? Where to draw the line? *Computer Law and Security Review*, 702-712.

Statistics How To, S. (2014). *Bayes Theorem: What is it used for in Statistics*. Retrieved March 23, 2014, from Statistics How To: http://www.statisticshowto.com/what-is-bayes-theorem/

Stenberg, D. (2014). *cURL*. Retrieved March 23, 2014, from cURL Manual: http://curl.haxx.se/docs/manpage.html

Varadharajan, V. (2010). Internet Filtering - Issues and Challenges. *Security & Privacy, IEEE, 8*(4).

Zellner, A. (2007). Generalizing The Standard Product Rule of Probability Theory and Bayes Theorem. *Journal of Econometrics, 138*(1), 14-23.