A New Simple Classification Algorithm enabling a New Approach for Identification of Virtual Bullying

K Burn-Thornton
University College
Durham University,
South Rd,
DURHAM DH1 3RW, UK.

T Burman
School of Engineering and Computer Science
Durham University,
South Rd,
DURHAM DH1 3LE, UK.

***Abstract*— In this paper we present a new, simple, classification algorithm which can be used to identify a change in virtual behaviour between a sender and recipient which could be used as an early indicator of virtual bullying or harassment. This application is not only, a novel application of Data Mining techniques but also, a new approach used to identify virtual bullying by virtue of identification of a change in behaviour.**

The approach which we have taken makes use of a new linear discriminant algorithm to classify normal and non-normal style(s) of email correspondence for each sender and recipient pair. A change in email style is taken to signify a change in relationship between the sender and recipient which could provide an early indicator of virtual harassment/bullying.

This approach has great potential for use in large organization where it is often appears to be hard to identify unacceptable information transmission between two colleagues – especially when one is in a more senior position.

By identifying behavior indicating a change in relationship between two colleagues it should be possible to instigate company anti bullying processes in a more timely manner and reduce the long term effect on those being bullied/harassed. This should ensure a more effective work force in terms of work place efficiency and reduction of stress related absence resulting from harassment or bullying.

We show that by regarding the contents of the emails as a set of Cascading Style Sheets, CSS, type files, which we call sender signature styles (SSSs), and accompanying information, it is possible to improve the identification of the number of sender signature styles contained within the email, irrespective of the length.

We also describe how, as a by-product of this work, a set of sender signature styles (SSSs) can be created during investigation of each email and hence be used as a library, containing increasing membership, for comparison with future emails sender by the same sender. By the nature of this task abnormal sender signature (ASSSs) files will also be created from virtual correspondence which has been known to be of concern, as well as that which has independently being identified as being indicative of being of possible concern

The implications of the use of SSSs, and ASSSs, for identification of future email interactions are discussed.

Keywords- Virtual Bullying, Data Mining, Novel approach.

I. INTRODUCTION

The current employment climate appears to have resulted in an increase in bullying and harassment experienced in the work place[1], This is indicated by the increase in the implementation of anti bullying/harassment procedures which have been put in place in company – as well as the proliferation of antibullying/harassment work place courses. This change in culture is also supported by many union web sites which proclaim on their site main page their success in fighting bullying cases which demonstrates its strong presence in a working environment [2].

Often this behavior is hard to identify, and eliminate, in large companies where such activity is readily facilitated by the virtual society in which this behaviour takes place by email and which can have a detrimental, and long-lasting, effect on those being bullied[3].

Approaches that have been taken to identify bullying behaviour and support those undergoing such behaviour include paper based 'tools' or process steps which require following[4].

Some software tools are also available but require virtual button pressing when bullying is taking place and are not an ideal approach for identification of one, or many, bulling or harassing emails. However, an approach which has not been taken is to make use of software to identify different nature, or construction, of emails which are sent from one sender to a recipient.

This paper describes a novel Data Mining approach, which enables a change in email style between a given sender recipient pair to be identified and hence provide a possible early indicator of virtual harassment/bullying by virtue of the change in virtual relationship.

The first section describes current approaches which are used to identify potential bullying/harassing behaviour in emails. This is followed by a discussion of two possible solutions which would enable email signature styles to be determined and a description of algorithms which may be gainfully employed in achieving each solution are then described.

An overview of the sub-tasks carried out by the algorithms which have been used to implement the proposed solution follows. The remaining sections discuss the investigations which were carried out in order to determine the effectiveness of the approach, the metrics which were used to determine the effectiveness and the results of the investigations for the CSS type solutions – the SSS(ASSS) based solution. Conclusions regarding the results of the investigations are then drawn with future profitable avenues for investigation being discussed.

II. EXISITING APPROACHES

The approach which is predominantly used in this area is that of the provision of reactive solutions when someone feels that they are being subject to cyber-bullying [5-6] and are not readily ideal to provide a solution to bullying/harassment from email.

These solutions range from papers based tools, or a series to steps to follow [4] , to software such as KnowDiss or CyberBully which are to all intents and purposes software in which the virtual pressing of a 'panic button' cause emails, or instant messaging, to be created to inform others what is happening or what is perceived to be happening [5-6].

In an ideal world a response to bullying behaviour should be proactive rather than reactive. If such an approach were to be taken the solution software would need to be able to identify bullying/harassing behavior as it were about to happen and not afterwards.

Such a proactive approach could be a solution which could detect a change in email style using a pattern matching approach based upon the fact that all emails have a unique signature style[2, 7, 8, 9](ASS) since all emails from the same sender should contain only one SSS or a variant on the same SSS in correspondence with the same recipient or group of recipients.

III. DOCUMENT (EMAIL) SIGNATURE STYLE

Document signature style makes the assumption that each individual has a unique writing style which is characterized by their individual use, and combination, of nouns, verbs and other features which include referencing[2, 8, 10, 11]. If the document signature style were to vary throughout the paragraphs of an email or between the sender and different recipients this could provide an indication that there was change in virtual behaviour between the sender and the recipient or recipients.

Such variation in style could be used as a basis for early instigation of any bullying/harassment in a company – something which is often hard to identify by the presence of hierarchical relationships and trans company support networks - especially if historical information could be used to show that early indications of a change in virtual behavior resulted in bullying/harassment behavior at a later date.

This approach could, if sufficiently accurate, prove to be a driver for facilitating a reduction in company sickness absence as well as progressing toward eliminating this unacceptable behaviour in a work place.

A Extraction of Signature Style

In order to determine the unique sender signature(s) present in the emails it is necessary to determine key elements of emails which can be used to determine a unique email signature created by each sender.

Initial analysis of over 1000 emails in one university School[9] suggested that the key elements of the signature required in order to determine whether, or not, an email is 'normal' or 'harassing/bullying' may be reduced to number of words in a sentence, number of lines in a paragraph, paragraph formatting, degree and use of grammar, type of language used and word spelling. These key signature features are concomitant with those proposed at ICADPR for instance those in [10] and [11].

The first two elements of the signature are self explanatory but the others may require some clarification. Degree and use of grammar to include the manner in which infinitives are used; use of, and types, of punctuation; use of pluarality. Type of language is taken to mean language style which in different types of English for instance UK and US. However, word spelling includes not only language spelling differences such as those found between UK & US, for example as in counsellor and counselor, but also frequency of typographical errors and spelling mistakes.

A solution to this analyzing this information would be an approach which is able to extract the key signature elements, and their values, from paragraphs, and compare them with others in the same email and with those extracted from other emails from the same sender. It could also be helpful if the approach used could be used, during any subsequent university formal procedures, to show how the email would have appeared if written in a 'non harassing/bullying' style by the sender. Such documentation would prove useful if additional proof 'virtual bullying', or change in virtual behaviour, towards a particular recipient, was required.

The following section describes two possible variants on such an approach.

IV. POSSIBLE APPROACHES

Both of the possible approaches suggested in this section make use of a modification of the approaches which we used in our web site maintainability tool[6] and multiple submission tool. The approaches make use of Cascading Style Sheets (CSS) or a combination of the eXtensible Markup Language (XML) in combination with the eXtensible Style Language (XSL) [13].

These approaches make use of information extraction and representation. Some commonality can be observed between the first steps of the approaches, which are described in the next section, and that of Ghani [14] and Simpson[15].

A CSS

If a CSS –based approach were used, a named sender signature style (SSS) could be defined which would describe the values assigned to the key signature features. Once the

SSS files were created, the signature of style of the sender could not only be compared with others within the same email but it could also be applied to any email section and the output compared with that contained within the current, or other, emails sent by the same sender. By using this approach the speed of investigation of emails could be minimized by the reduction in the size of file which is required in order achieve comparison [16].

In practice, each section of the email being investigated could be converted directly to a section of SSS containing the feature values. Such an approach would require the use of a measure of uncertainty when mapping the samples of document and related SSS code to named signature styles. Figure 1 provides an example of how a page of email text may be converted using such an approach.

Data Mining would appear to be able to provide a solution to this problem by making use of modified clustering techniques.

The only drawback to this approach is that a library of assignable values for each key signature feature will need to be defined initially. However, this library could be updated as part on an electronic backup process.

B XML

For an XML approach all content information would be contained in an XSL file with its companion XML file containing the ASS feature information which would be recursively applied to the XSL document.

Using the example from Figure 1 this approach would result in the production of a XML file containing a section of text that would be marked up as a reference name, and the XSL file would contain a template which could be applied reference names in that document. Such an approach would readily facilitate comparison of emails because it would be relatively easy to target comparison of emails by investigation of specific signatures, SSSs.

Rigid definitions do not exist for XML tags which means that any appropriately defined names will have to be used in the XML file as well as a library of attributable values of the signature features, as in the CSS approach. However, a major drawback of this approach would be the need of consistency for XML tags and the possibility of ongoing modification to a centrally accessed XML tag dictionary.

The requirements which will need to be fulfilled for the XML/XSL solution suggest that the CSS based solution may be the more accurate approach to use for the comparison of signature styles in emails. This is because even a slight variation is XML tags could result in a large discrepancy in ASS and hence identification of a document as containing more than one sender style when it does not.

The following section provides an introduction to Data Mining, which will be used as the basis of the CSS , or SSS, approach.

V. *SUITABLE DATA MINING APPROACHES*

The class of algorithms, or approach which we can utilize more effectively, appears to be from the statistical class of algorithms.

These are the same algorithms which were discussed for the task of web site maintainability [6]. The reasons behind the choice of algorithm for the task are discussed in the final sub-section.

The most appropriate algorithm for the conversion from sender email to CSS, SSS, from those listed above, is the k-NN algorithm, or a variant of such. The other algorithms are not appropriate because they either require too many samples with which to build an effective model from which to work effectively in this application (decision trees, Bayesian classifiers), require numerical data (Fisher's linear discriminants), or require prior knowledge of the classes (K Means).

However, k-NN can work effectively with a small number of samples, can work with categorical data given an appropriate function to compare two samples, and does not require any prior knowledge of the number of classes, or sender types.

The following sections describe the implementation of the CSS solution which has been described in this section.

VI. CSS SOLUTION

In order to implement the k-NN algorithm, or a slight variant therein, some means of finding a numeric difference between two samples of the senders emails and SSS is required. This can be achieved by determining the percentage of signature features in one sample which are not present in the other sample or samples.

A visual representation of the approach used to determine the difference between the two samples of SSS signature features, and their values, present in each sample signature, may be seen in Figure 2.

In order to achieve this each section of email needs to be represented by equivalent signature features and their values. In the same manner as presentation tags in HTML code these can be represented as signature tags. It is these adjacent signature tags which form clusters of tags and can be represented by a single SSS.

The first stage of the implementation of the k-NN type algorithm, kb-NN, is to create the signature tags from the original document and then each cluster of signature tags is converted to a SSS sample using a set of rules that are defined in a data file. This can be changed by the user as the SSS evolves, but a standard set of rules.

Each line is in the format:

Tag-name SSS-equivalent value

After each cluster is converted to an SSS the algorithm iterates through each sample and compares it to any that have already been classified. At the start of the loop, none will have been classified. Otherwise, a list of the other classified samples is created and ordered by difference to the new sample. If no sample is within a threshold distance, it is assumed that the new sample is not sufficiently similar to any previous classification, and so the user is prompted for a new classification for this sample. Otherwise, the closest k

samples are taken from this list and the new sample is assigned the same classification as the majority of these k samples. An appropriate value of k can be found through trial and error during initial investigations.

For the final conversion of the classifications to a style sheet, an arbitrary sample from each classification is used to supply the definition of the style, and the name assigned to the classification is used as the name of the style. As each sample in the class should be very similar, it should not matter which sample is used for the style definition.

A slight modification was made to the kb-NN class so that it could be used to create an example document from an existing signature style. This modification was that a new sender signature is not created if no close match among the previously classified samples is found i.e. if a change in email style exists in the email. The contents of the style sheet are read in and set as the classified samples to provide the classification.

The same approach is used for finding groups of email paragraphs with the same style. The major differences in this case is that the methods used to represent each paragraph, and the differences between them – as well as the automatic naming procedure of a process which is to all intents and purposes completely unsupervised.

Each paragraph, is represented by a set of feature information, including a list of the number of times each one is used, and the distribution of the feature tags throughout the page or paragraph. The combination of this set of information gives a good overall impression of the written signature style of the sender.

The difference between two sets of information is found by the number of features, and values, that are not present in one set of information and is present in another, or those where the font is used more than twice as many times in one than in the other. The table distributions are compared using the chi-squared test. Each distribution is composed of 100 values, indicating the number of signature tags in that 100th of the section. The chi-squared value is calculated as the sum of the squares of the differences of each of these values, as given by the formula:

$$\chi^2 = \sum_{i=1}^{100} \frac{(x_i - y_i)^2}{y_i} \qquad \text{[equation 1]}$$

where

x is distribution of table tags in information1.
y is distribution of table tags in information2.

The set of this information provides an overall value for the difference between the two emails, or paragraphs.. This can then be directly compared to the value for any other emails. Again, if the email. or paragraph, being classified is not sufficiently similar to any previously classified section, a new classification, or SSS, is created for it.

The following section describes investigations which were carried out, using the new algorithm, to determine the effectiveness of the CSS methods to facilitate comparison of sender signature styles (SSS) in emails.

VII Investigations

In order to determine the effectiveness of the approach used, a set of metrics were defined which enabled the effectives of the solution to be determined on a wide range of emails sent.. This section describes the metrics used and the wide range of emails used.

A Measures of Effectiveness: Metrics Used

The effectiveness of the solution was determined by the ease, and effectiveness, of extraction of file information from the source email into a separate sender signature style sheet and the degree to which the content of the original emails remained unaltered once it has been produced by use of the style sheet.

The metrics of :- Number of sender signature styles produced and number of differences between the sender style features in the original email, or paragraph, in the email and that created using the SSS were also used to determine the effectiveness of the solution.

The following sections describe in detail the metrics and provides a justification for their use.

Metric 1 - A count of the sender (email) signature styles produced.

Sections of email paragraphs which are slightly different could potentially be converted to the same SSS style, because the data mining approach used allows for some fuzziness in the classification in line with a sender styles varying slightly within the paragraphs of an email. However, email paragraphs which vary greater than observed with one email should result in different SSS styles. This should be indicated by the number of styles produced. Therefore the number of styles produced is also an important measure of how easy it will be to determine commonality in sender signature style within paragraphs contained within an email.

Metric 2 – Information Differences

The key sender signature features emails created by the system, using the appropriate SSS, should identical to those contained within the original, or other, emails. This is tested by measuring the number of differences between the original and newly produced emails, assigning a score to each type of difference, and adding these scores together.

B Emails Investigated

Figure 3 provides examples of the wide range of emails which were investigated.

These emails were chosen as examples of their wide range of emails to which the new algorithms can be applied because they represent a cross section of the variation in sender styles contained with emails sent within a university school.

Sample 1 containing emails sent by an email sender who has never been known to be the subject of a complaint regarding harassment/bullying.

Sample 2 contains emails sent by a sender whose first language is not English and containing emails sent by an email sender who has never been known to be the subject of a complaint regarding harassment/bullying.

Sample 3 containing emails sent by an email sender who has been known to be the subject of a complaint regarding harassment/bullying.

This range of emails should enable the performances of the new algorithm on different styles of emails to be determined.

The following section describes the results from applying the metrics to the wide range of test emails.

VIII. RESULTS

Simple plots are used to visualize the results. Figures 4 to 5 show the results of investigation of the two metrics.

A Count of the sender signature styles produced.

The number of sender styles produced is dependent of the written content of each email. Figure 4 shows that, on average, two styles are produced from a email known to have one sender signature style. The figure also shows that, on average, three styles are produced from an email of unknown type with the distribution of the number of styles produced being skewed towards the lower end. The new algorithm accurately determined the number of the sender email types from the emails known to be of bullying/harassing nature However, the figure shows that human determination was less accurate – especially for samples 2 those for which English was not a first language.

B Information Differences

These results shown in Figure 5 are consistent with that results of the SSS investigations in that information differences observed between the original, and key features of the, email are strongly correlated with the error in determining sender type. Thus suggesting that if the SSSs contained in the email can be determined then it is it possible to reform key features of the original email for comparison with other sender emails and with future emails by the same sender.

IX. CONCLUSIONS & FUTURE WORK

We have described a novel application of Data Mining in which a new linear discriminant algorithm, kb-NN, a variant on k-NN, , which enables an indicator of a change in virtual relationship between the sender and recipient, an hence an early indicator of possible virtual whether emails sent are of a bullying/harassment

The results presented in section VII show that the approach used facilitates accurate investigation of the nature of emails send by a specific sender and indicate whether virtual bullying/harassment may be occurring. Such results have the potential to be used in early instigation of anti harassment/bullying procedures..

Is intended that further work will be carried out investigating the three key metrics in email from other Faculties and universities. Work will also be carried out to modify the Data mining algorithm to maintain accuracy of indication of potential bullying/harassing emails across this new range of email documents.

ACKNOWLEDGEMENTS

Acknowledgement is made to Mark Carrington for his original project work in 2002 which led to development of this paper.

X. REFERENCES

[1]www.bohrf.org.uk/downloads/bullyrpt.pdf, accessed 18/4/2012.

[2] www.ucu.org.uk/media/pdf/f/0/bully_harass_toolkit.pdf accessed 17/4/2012.

[3] http://www.Mind.org.uk accessed 18/4/12.

[4]www.nhs.uk/Livewell/workplacehealth/Pages/bullyingat work accessed 18/4/12.

[5] news.com.au, April 19, 2011 accessed 18/4/12.

[6] http://www.KnowDiss.com accessed 18/4/12.

[7] Cai J, Paige R and Tarjan R, More Efficient Bottom-Up Multi-Pattern Matching in Trees, Theoretical Computer Science, 106), pp.21-60,1992.

[8] Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 1 Writer Identification in Handwritten Documents Curitiba, Parana, Brazil September 23-September 26ISBN: 0-7695-2822-8

[9] Information produced from Brunel University under FOI **Act.**

[10] Chaski, C. E. , 2007-07-25 "Multilingual Forensic Author Identification through N-Gram Analysis" Paper presented at the annual meeting of the The Law and Society Association, TBA, Berlin, Germany 2010-06-04 from http://www.allacademic.com/meta/p177064_index.html.

[11] Siddiqi I, Vincent N, "Writer Identification in Handwritten Documents," Document Analysis and Recognition, International Conference on, vol. 1, pp. 108-112, Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 1, 2007.

[12] Kövesi B, Boucher JM, and Saoudi S, Stochastic K-means algorithm for vector quantization. Pattern Recognition Letters, 22,pp. 603-610, 2001.

[13] Wilde E, Wilde's WWW. Technical foundations of the World Wide Web. London: Springer, 1999.

[14] Ghani R, Jones R, Mladenic D, Nigam K and Slattery S, Data mining onsymbolic knowledge extracted from the web, in Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000), Workshop on Text Mining.

[15]Simpson S http://www.comp.lancs.ac.ucomputing/users/ss/websitemgmt , accessed 10/2/12.

[16] Sommerville I, Software engineering 5th ed., International computer science series, Wokingham, England : Addison-Wesley, 1996.

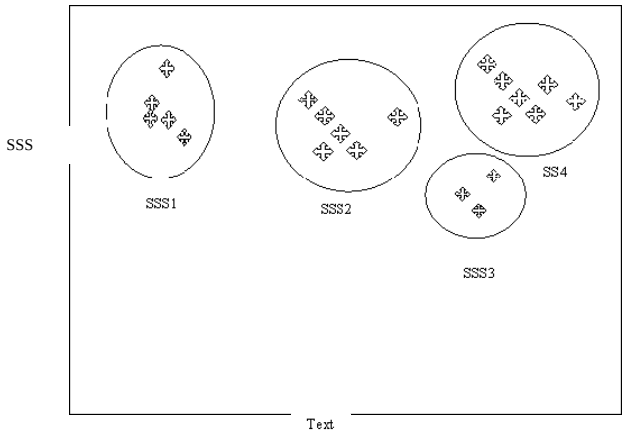

Figure 1- Clustering

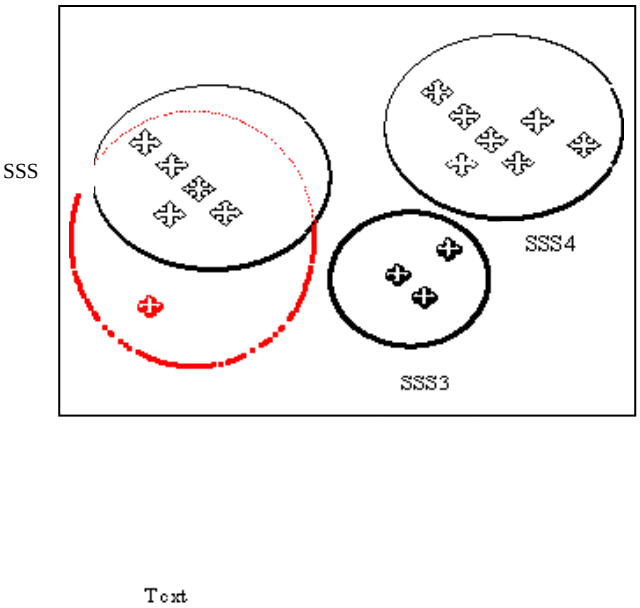

Figure 2 – kb-NN Classifying email

Sample	Staff	First Language	Number	Email Type
1	UK	English	100	Known 'Non Bullying/Harassing'
2	UK	Not English	100	Known 'Non Bullying/Harassing'
3	UK	English	100	Known ' Bullying/Harassing'

Figure 3 - Examples of email types.

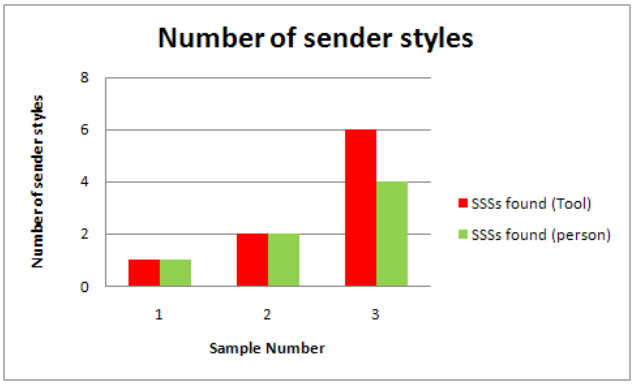

Figure 4 - A count of the sender signature styles produced.

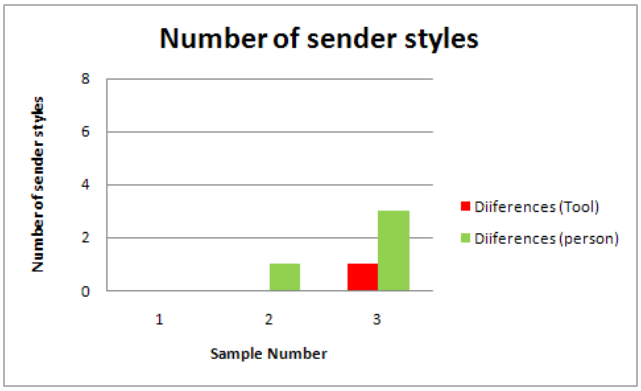

Figure 5- Information Differences between Original and Reformed Email