Improved cyberbully detection techniques using multiple correlation coefficient from forum corpus

J.I. Sheeba*

Department of Computer Science and Engineering,
Pondicherry Engineering College,
Puducherry, India
Email: sheeba@pec.edu
*Corresponding author

S. Pradeep Devaneyan

School of Mechanical and Building Sciences,
Christ College of Engineering and Technology,
Puducherry, India
Email: pr.signs@gmail.com

Prathyusha Tata

Sapient Corporation,
Publicis.Sapient,
Bangalore, India
Email: prathyusha@pec.edu

Abstract: Today, there are many prominent online sites where people share their experiences regarding crimes and anti-social behaviour. In this regard, a major unaddressed and even unidentified problem that is experienced in the social network websites is *cyberbully*. This proposed framework primarily targets the cyberbullying in the crime investigation forum since a high degree of cyberbully is common in crime forums. In this paper, a highly furnished representational framework is proposed that is specific to cyberbully detection using hybrid techniques (multiple correlation coefficient – MCC and support vector machine – SVM). The bag of words are given individual weights to examine their correlations using MCC algorithm before feeding them into a linear SVM classifier that identifies and classifies the cyberbully words. The efficiency of the system developed can be enhanced by analysing the evaluation metrics and the dataset validation metrics.

Keywords: cyberbully detection; cyberbully classification; multiple correlation coefficients; MCCs; support vector machines; SVMs; data analytics; DAs.

Reference to this paper should be made as follows: Sheeba, J.I., Devaneyan, S.P. and Tata, P. (2018) 'Improved cyberbully detection techniques using multiple correlation coefficient from forum corpus', *Int. J. Autonomic Computing*, Vol. 3, No. 2, pp.152–171.

Biographical notes: J.I. Sheeba received her BE in Computer Science and Engineering from the Bharathidasan University and ME in Computer Science and Engineering from the Anna University. She has completed her PhD in the

area of text mining. She is presently working as an Assistant Professor from the Department of Computer Science and Engineering, Pondicherry Engineering College and has more than 14 years of teaching experience. She has published papers in many international journals and conferences. Her research interest includes data mining, soft computing and information security.

S. Pradeep Devaneyan has completed his BE in Mechanical Engineering from the Karunya Institute of Technology, ME in CAD/CAM from the Arulmigu Kalasalingam College of Enginering and Technology and PhD from the Pondicherry University. He is currently working as a Professor at the Christ College of Engineering and Technology, Puducherry and has 14 years of teaching experience. He has published papers in many reputed journals and conferences.

Prathyusha Tata received her BTech in Computer Science and Engineering from the Pondicherry Engineering College. She is currently working as an Associate Technology L1 at the Publicis.Sapient. Her areas of interest are text mining and data analytics.

This paper is a revised and expanded version of a paper entitled 'Cyberbully detection using hybrid techniques' presented at International Conference on Telecommunication, Power Analysis and Computing Techniques (ICTPACT 2017), Chennai, India, 6–8 April 2017.

1 Introduction

The number of people using social networking sites has greatly increased since the last few years. The users are given privilege to create their own profiles and to communicate with other users without much complicated terms and conditions. These sites are becoming the sources of large dynamic data. This even lead to the popularity of cybercrimes like phishing, spread of malware and cyberbullying (Al-garadi et al., 2016). In particular, cyberbullying has emerged as a major problem along with the recent development of online communication and social media. Cyberbullying has also been extensively recognised as a serious national health problem, in which victims demonstrate a significantly high risk of suicidal ideation. One study conducted by national anti-bullying charity Ditch the Label in 2013, has shown that two out of three 13–22 years old who were surveyed have been victims of cyberbullying (Zhao et al., 2016). As reported in Ybarra (2010), approximately 43% of teens once reported being bullied through social media. Another study also shows that cyberbullying victimisation rate ranges from 10% to 40% (Kowalski et al., 2014). A cyberbully (a person involved in cyberbullying) can annoy his/her victims before an entire community, which may result in the negative effects on the victim. In this regard, cyberbullying is to be controlled and informed to the authorities. Accordingly, the posts that include the cyberbully words (such as insulting, threatening, terrorism, bad, vulgar phrases) are recognised.

Unfortunately, cyberbullying is most of the times unidentified and unaddressed. And even worse is that the bullying messages are left on the internet creating much more issues to the other users. To detect cyberbullying content underlying huge volumes of posts on social media, a good solution is to develop machine learning-based automatic cyberbullying detection system, so that all the sensitive information would be modified or

erased at the first time, preventing internet users from overexposure to undesirable information.

The basic functioning of the system is to find the most relevant and important words or features in the large dataset. High accuracy can be obtained by working on the reduced sets. The proposed system calculates the correlation between the existing words with the words identified from the input. To avoid data redundancy, the data with high value of correlation are alone considered. The proposed framework involves reducing the redundancy in the dataset using data pre-processing techniques and identifying the most related words, i.e., feature extraction using multiple correlation coefficient (MCC) (Bilski, 2014). A linear support vector machine (SVM) classifier is used to identify the class label of each word in the dataset. The identified word is classified in any of the classes insulting cyberbully, vulgar cyberbully, threatening cyberbully, bad cyberbully and terrorism cyberbully (Dadvar et al., 2012).

Data analytics (DAs) is the science of analysing data to filter useful knowledge from the information. This knowledge could help us understand our world better, and in many contexts enable us to make better decisions. With this objective, the last 20 years has seen steeply decreasing costs to process data, creating a stronger motivation for the use of statistical approaches to problem solving. This paper presents a wide range of DA techniques and is structured around the broad contours of the different types of DAs, namely, descriptive, inferential, predictive, and prescriptive analytics.

SAS stands for *Statistical Analysis Software*. It was developed in the year 1960 by the SAS Institute. From 1st January 1960, SAS was used for data management, business intelligence, predictive analysis, descriptive and prescriptive analysis, etc. With the introduction of JMP (jump) for statistics SAS took advantage of the graphical user interface which was introduced by the Macintosh. JMP is basically used for the applications like Six Sigma, designs, quality control and engineering and scientific analysis. SAS is platform independent which means you can run SAS on any operating system either Linux or Windows. Over the years SAS has added numerous solutions to its product portfolio. It has solution for data governance, data quality, big DAs, text mining, fraud management, health science, etc. We can safely assume SAS has a solution for every business and financial domain (Prathyusha et al., 2017).

The main objective of this paper is to identify the cyberbully words (such as insulting, threatening, bad, vulgar, terrorism words) in the crime investigation forum using hybrid techniques. This paper is organised in such a way that in Section 2, initially the related works are introduced and then the proposed machine learning model for cyberbully detection is presented in Section 3. In Section 4, experimental results on a real time forum corpus are illustrated and analysed. Finally, concluding remarks and future works are provided in Section 5.

2 Related works

Al-garadi et al. (2016) proposed a model that provides a feasible solution to detecting cyberbullying in online communication environments. This is an approach for offensive language detection that was equipped with only a simple lexical syntactic feature. Zhao et al. (2016) proposed a representation learning framework specific to cyberbullying detection. The words are assigned with different weights to obtain bullying features,

which are then concatenated with the bag-of-words and latent semantic features to form the final representation. Here, based on word embeddings, a list of pre-defined insulting words is given which made the system precise. Bilski (2014) presented the application of statistical (econometrics-originated) methods to process datasets used by the artificial intelligence techniques in the diagnostics of analogue systems. Before the training and evaluation of the intelligent modules is performed, the measurement data are analysed to minimise the number of attributes (symptoms) required to distinguish between different states of the system under test (SUT). The pre-processing increased complexity exponentially.

Abdelhaq et al. (2016) describes the keyword extraction for localised events in twitter. The algorithmic steps involve building keyword geo reference map, locality check and focus estimation, topical to local keyword recovery. The work was especially to extract local keywords, i.e., busty words that have a very limited spatial extent. Orencik et al. (2016) explained a privacy-preserving search over encrypted data using queries with multiple keywords. The algorithmic steps involve index generation, query generation, two-server secure search and document retrieval over the cloud data. The paper is described for a complex public storage system model. An SVM model for training text classifier which is gender-specific and to detect the written language by a harasser including gender which varies with feature is proposed by Dadvar et al. (2012). In this, gender alone is taken into consideration which is not enough for complete detection. Kowalski et al. (2009) studied about electronic bullying among middle school students and the results showed that 84% of school students have experienced bullying in this study.

Chen et al. (2012) developed an approach for bullying detection that was equipped with a lexical syntactic feature, although lexical features perform well in detecting offensive entities without considering the syntactical structure of the entire sentence, they fail to distinguish sentence offensiveness which contain same words but in different orders. McGhee et al. (2011) proposed computer software to detect the presence of cyberbullying in the online chat conversations. People with knowledge about software alone can make use of it. Lossio-Ventura et al. (2016) proposed a methodology that offers several measures based on linguistic, statistical, graphic and web aspects. These measures extract and rank candidate terms with precision results for automatic term extraction, and work with different languages. The system lacks in frequency information with linguistic pattern.

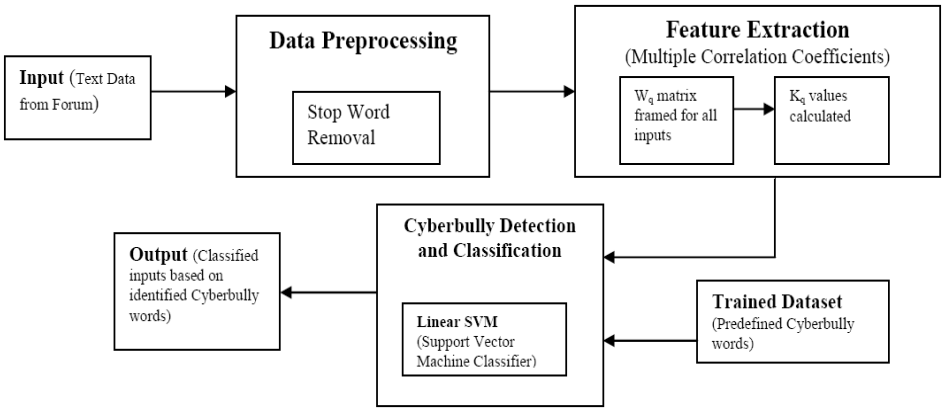
Pradheep et al. (2018) introduced a method for detecting multimodel cyberbullying such as audio, video, image along with text in the social networks. The cyberbully images are detected using the computer vision algorithm and the cyberbully videos are detected using the shot boundary detection algorithm. Sheeba and Devaneyan (2018) proposed a new technique which detects cyberbullying from direct, indirect cyberbullying techniques and to prevent the cyberbullying by reflective messages and also to recover the victims using depression technique. Rafiq et al. (2018) proposed a multi-stage cyberbullying detection solution that drastically reduces the classification time and the time to raise alerts. This system is highly scalable without sacrificing accuracy and highly responsive in raising alerts. Raisi and Huang (2018) introduced a method for detecting harassment-based cyberbullying using weak supervision. In this framework in which two learners train each other to form a consensus on whether the social interaction is bullying by incorporating nonlinear embedding models.

Most of the existing systems lack in an effective use of classification systems. To overcome these problems, in this proposed framework, the correlations among the symptoms are identified as an additional feature that reduces the SUT before classification is handled.

3 Proposed framework

The working of the proposed system, i.e., the process of detecting cyberbully words from the given input dataset is described detail in Figure 1. The dataset is taken from the conversations of the users of 4forums.com. To improve the quality of the research data, the dataset is pre-processed using a technique namely 'stop word removal'. To reduce the computational complexity the most important features are obtained using the feature extraction techniques. The words weights are calculated and are sent to feature extraction module. The inputs that result with a high correlation value that is calculated using the MCC algorithm is screened out from the whole dataset and is sent to the classifier module. The various cyber bully words present in each input are identified and are classified into insulting cyberbully, vulgar cyberbully, threatening cyberbully, bad cyberbully and terrorism cyberbully using the linear SVM classifier.

Figure 1 Proposed framework for *cyberbully detection and classification system using MCC and SVM* (*CDMS*)



The working of the system can be explained detail in three steps. They include data pre-processing, feature extraction and cyberbully detection and classification.

3.1 Data pre-processing

Data pre-processing methods are required to reduce the huge set of irrelevant and redundant data from the forum data. Forum data is usually too noisy. The use of data pre-processing methods reduces the number of unnecessary comparisons that the system may incur and this in turn reduces the execution time. The data pre-processing is executed using the process of stop word removal.

Stop words are words which are filtered out before or after processing of natural language data. Stop words refer to the most common words in a language. These are some of the short function words such as is, the, at, which and on. Stop word removal is important to reduce the data redundancy and irrelevancy.

3.2 Feature extraction

Feature extraction starts from an initial set of measured data and builds derived values intended to be informative and non-redundant. When the input data to an algorithm is too large to be processed and it is suspected to be redundant then it can be transformed into a reduced set of features. The process of feature extraction is handled using the algorithm MCC.

MCCs are a measure of the strength of the association between the independent and explanatory variables and the one dependent (predictor) variable. It selects the combination of the symptoms correlated strongest with the faulty parameter. The original set is divided into sub-sets, each containing few attributes of the faulty parameters. Each set is then processed separately.

Consider the whole dataset to be D_0, the separate simplified datasets obtained after the stop word removal is D_1.

3.2.1 Finding the most important set of symptoms

After removing the stop words, only a simplified dataset is left with a reduced number of words. From them, the optimal sets of features that are more related to the training data are retrieved. This selection results in a cardinality of $2^p - 1$ number of combinations.

The data structures used by this method are:

$$R_0 = \left[r_1, \ldots\ldots, r's\right]^T \tag{1}$$

wherein equation (1), R_0 is the vector of words in the training dataset that is a set of predefined cyberbully words.

$$r_j = \frac{\sum_i \left(s_{ij} - \widehat{s_j}\right) \cdot \left(c_i - \hat{c}\right)}{\sqrt{\sum_i \left(s_{ij} - \widehat{s_i}\right)^2 \cdot \sum_i \left(c_i - \hat{c}\right)^2}} \tag{2}$$

Here in equation (2), s_{ij} is the value of the j^{th} word in the i^{th} row, $\widehat{s_i}$ is its mean value, c_i is the value of the identified cyberbully word in the i^{th} row, and $\hat{c}$ is its mean value.

The second structure is the similarity matrix between pairs of words from the input (i^{th} and j^{th}), with the analogous interpretation of the similarity values.

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & 1 & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{3}$$

The MCC method uses these two data structures to calculate the MCCs K_q for each q^{th} combination of input.

$$K_q = \sqrt{1 - \frac{\det(W_q)}{\det(R_q)}} \tag{4}$$

wherein equation (4), $\det(R_q)$ is the determinant of the similarity matrix R in equation (3), i.e., the similarity value for the words present in the q^{th} combination and $\det(W_q)$ in equation (4) is the determinant of the matrix constructed from the correlation vector R_0 in equation (1) and the correlation matrix R in equation (3).

$$W_q = \begin{bmatrix} 1 & R_{0q}^T \\ R_{0q} & R_q \end{bmatrix} \tag{5}$$

The algorithm is implemented for each of the inputs from the entire dataset. The input with the greatest value of K_q [equation (5)] is identified as the cyberbully and is given to the next module. The main disadvantage is the high computational complexity, determined by the number of words' combinations: $O(2^p - 1)$. This makes the stop word removal the crucial stage (Bilski, 2014).

Thus the output obtained from the MCC algorithm is a set of inputs that are most related to the cyberbully. This is the part of the dataset the system is actually interested in. The output obtained at this stage is given as input to the cyberbully detection and classification module where the cyberbully words present in the input are identified and a class label is assigned.

3.3 Cyberbully word detection and classification using SVM

The reduced and refined dataset is sent to a linear SVM classifier. The given input dataset is cross examined against a training data. The training data is given as a diversified collection of cyberbully words, i.e., insulting, bad, vulgar, threatening and terrorism words. In machine learning, vector machines act as supervised learning models with associated learning algorithms that analyse data used for classification. Given a set of training symptoms, each one marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new symptoms to one category or the other, making it a non-probabilistic binary linear classifier. The SVM model is a representation of the symptoms as points in space, mapped so that the symptoms of the separate categories are divided by a clear gap that is as wide as possible. New symptoms are now mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

More formally, a SVM constructs a hyperplane or some set of hyperplanes in a high or infinite-dimensional space, which can be used for classification. Intuitively, a good separation is attained by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the generalisation error of the classifier is inversely proportional to the margin.

Wherein, the original problem may be stated in a finite dimensional space, it is most common that the sets to discriminate are not linearly separable in that space. For this reason, it is proposed that the original finite-dimensional space be mapped into a higher-dimensional space, making the separation in that space easier. To keep the

computational complexity reasonable, the mappings used by these schemes are designed in a way that these dot products may be computed easily in related terms of the variables in the original space, by defining them in terms of a kernel function $k(x, y)$ that is defined to suit the problem. These selected hyperplanes in the higher-dimensional space are defined as a set of points whose dot product with a vector in that space always remains constant. The vectors that define the hyperplanes can be chosen to be linear combinations with parameters f images of feature vector database x_i that occur in the database. With this choice of a suitable hyperplane, the points x in the feature space that are mapped into the hyperplane are defined by the following relation:

$$\sum_i \alpha_i k\left(x_i, x\right) = \text{constant} \tag{6}$$

Note that if $k(x, y)$ becomes small as y grows farther away from x, each term in the sum measures the degree of closeness of the test point x to the corresponding database base point x_i. In this way, the sum of all the kernels can be used to measure the relative nearness of each test point to the data points. Note the fact that the set of points x mapped into any of the hyperplanes can be quite concluded as a result, allowing much more complex discrimination between sets which are not convex at all in the original space (Dadvar et al., 2012).

The linear SVM classifier gives the identified cyberbully words into five classes based on the given training dataset. The output is obtained as the classification of the identified cyberbully words along with the text classification based on the input and training dataset (Prathyusha et al., 2017).

4 Results and discussion

4.1 Dataset description

The data that is being analysed is the *forum dataset*. A forum (news group) is a place where people gather to discuss various topics and subjects. People can also discuss a common topic and share information with one another. The forum described here is http://www.4forums.com. This forum contains around 30 topics and five formal debates (such as techniques, styles, tournaments, open debate, comments). Particularly, crime debate forum dataset contains 198 threads and 7,621 posts.

4.2 Existing system

4.2.1 Cyberbully detection and classification system using GenLeven algorithm and fuzzy classifier

This framework is proposed to detect the cyberbully content present in the social network using GenLeven algorithm and to classify the detected cyberbully content as harassment cyberbully, insult cyberbully, terrorism cyberbully or flaming cyberbully using fuzzy rule base. The input is textual conversation which is pre-processed by removing stop words and then feature extraction is performed. The feature extraction involves extracting noun, adjective and pronoun words from the input and their corresponding frequency. GenLeven algorithm extracts the cyberbully words from the conversation using trained

dataset. The cyberbully words detected is classified using fuzzy classifier which performs classification using fuzzy if-then rules (Sheeba and Devaneyan, 2016).

4.3 Metrics considered for evaluation

The performance of the proposed framework was measured in terms of the quality measures, namely recall, precision, F-measure, root mean square error (RMSE), classification accuracy, sensitivity and specificity.

4.3.1 F-measure

F-measure computes both precision and recall. It can be estimated by using the given equation:

$$\text{F=2.}\frac{precison.recall}{precison + recall} \tag{7}$$

4.3.2 Classification accuracy

Accuracy calculates the proportion of correctly identified cyberbully words, and it is estimated by using the following equation:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{8}$$

In respect of cyberbully detection the terms are evaluated in the following manner:

TP determined as a word being classified correctly as relating to a cyberbully category

TN determined the words which were non-cyberbully words

FP determined as a cyberbully word even if it is in the non-cyberbully category

FN determined as a non-cyberbully word even if it is in cyberbully category.

4.3.3 Root mean square error

It is the difference between classifying the cyberbully words predicted by a system and the cyberbully words actually observed from the input. It is estimated by using the following equation:

$$\text{RMSE} = \sqrt{\frac{\sum_{i-1}^{n}\left(X_{obs,i} - X_{model,i}\right)^2}{n}} \tag{9}$$

where X_{obs} is the manually classified cyberbully words, X_{model} is the system classified cyberbully words at time/place i and n is the number of inputs.

4.3.4 Sensitivity (also called true positive rate)

Sensitivity is ability to identify a condition correctly. It will classify cyberbully words which are under the cyberbully words category in the given input. It is estimated by using the following equation:

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \tag{10}$$

4.3.5 Specificity (also called true negative rate)

Specificity is ability to exclude a condition correctly. It will classify cyberbully words which are not under the cyberbully words category in the given input. It is estimated by using the following equation:

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \tag{11}$$

Figure 2 shows the difference in F-measure gained by two different techniques. CDMS denotes cyberbully detection using MCC and SVM and CDCSGF denote cyberbully detection and classification system using GenLeven algorithm and fuzzy classifier. The evaluation of the system is carried on the dataset taken from 4forums.com. Figure 2 shows better F-measure values of CDMS than CDCSGF. Figure 3 shows the higher accuracy of CDMS over CDCSGF. Figure 4 shows the RMSE graph in which system using CDMS shows less error value than a system using CDCSGF technique. Sensitivity and Specificity values are calculated on randomly shuffled 50 conversations and the results obtained are shown in Figure 5 and Figure 6. It is inferred from Figures 5 and 6 that CDMS has achieved the best sensitivity and specificity measure when compared with existing technique CDCSGF.

Figure 2 F-measure (see online version for colours)

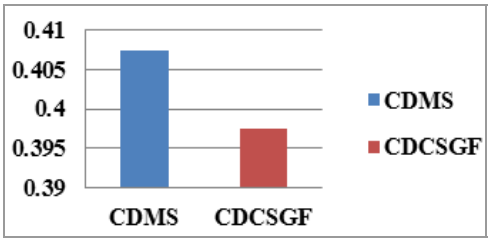


Figure 3 Classification accuracy graph (see online version for colours)

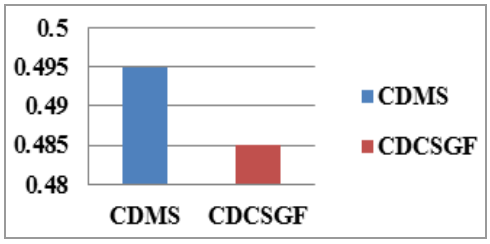


Figure 4 RMSE graph (see online version for colours)

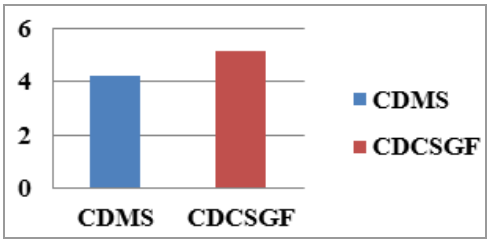


Figure 5 Sensitivity graph (see online version for colours)

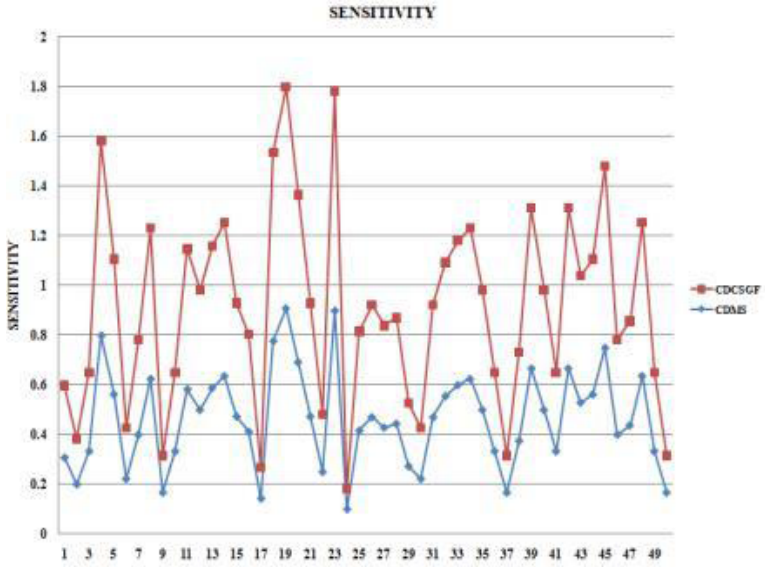
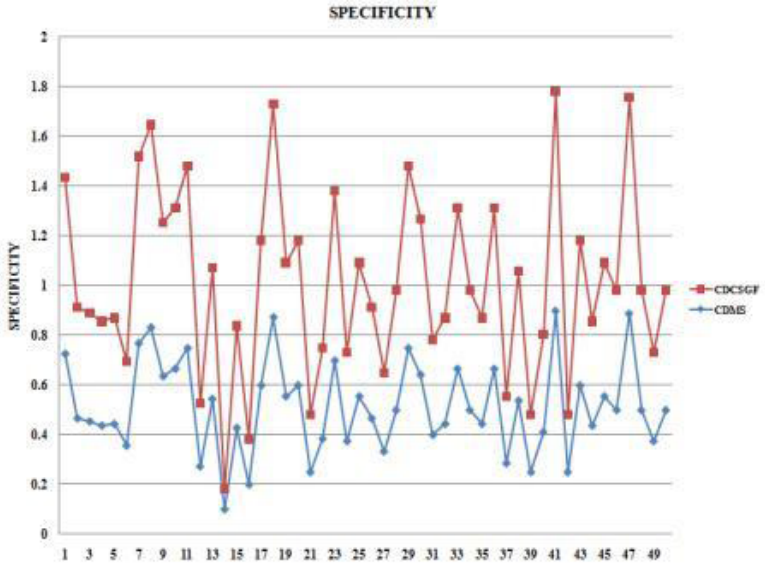


Figure 6 Specificity graph (see online version for colours)



4.4 DAs using SAS

DA is the process of examining datasets in order to draw conclusions about the information they contain. Data is extracted and categorised to identify and analyse behavioural data and patterns. DAs are the pursuit of extracting meaning from raw data.

4.4.1 Cross tabulations

Cross tabulation involves producing cross tables also called contingent tables using all possible combinations of two or more inputs. For example, the contingent tables for a sample of data from the dataset are drawn and the frequency distributions are calculated.

Table 1 displays the frequency of each word in the samples _1 and _2. The same procedure is repeated for all possible combinations of the inputs in the dataset. The row total and column total indicates the frequency of the combination of the words.

Table 1 Cross tabulation

Table of _1 by _2															
_1	*_2*														
Frequency	*11*	*25*	*31*	*42*	*43*	*49*	*61*	*63*	*65*	*75*	*81*	*87*	*93*	*95*	*Total*
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
30	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
32	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
58	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
80	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
87	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
92	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
95	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
96	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Total	1	0	1	1	1	1	0	1	0	1	0	1	0	1	9
Frequency missing = 5															

4.4.2 T-tests

The T-tests are performed to compute the confidence limits for one sample or two independent samples by comparing their means and mean differences. The paired value sets are determined based on the class labels and the inputs.

Table 2 shows the values drawn from the inputs like mean and standard deviation. These are the identical values of the samples.

Table 2 T-test

N	*Mean*	*Std. dev*	*Std. err.*	*Minimum*	*Maximum*
30	50.3667	27.4735	5.0159	3.0000	95.0000

Considering the confidence level to be 95%, the mean and standard deviation are calculated and the values are shown in Table 3.

Table 3 Standard deviation

Mean	*95% CL mean*		*Std. dev.*	*95% CL std. dev.*	
50.3667	40.1079	60.6254	27.4735	21.8801	36.9330

The result of the t test that is the t value and the degrees of freedom (DF) values are presented in Table 4.

Table 4 T-value

DF	*t-value*	*Pr* > |*t*|
29	10.04	< .0001

Figure 7 Distribution of the input for mean with 95% confidence interval (see online version for colours)

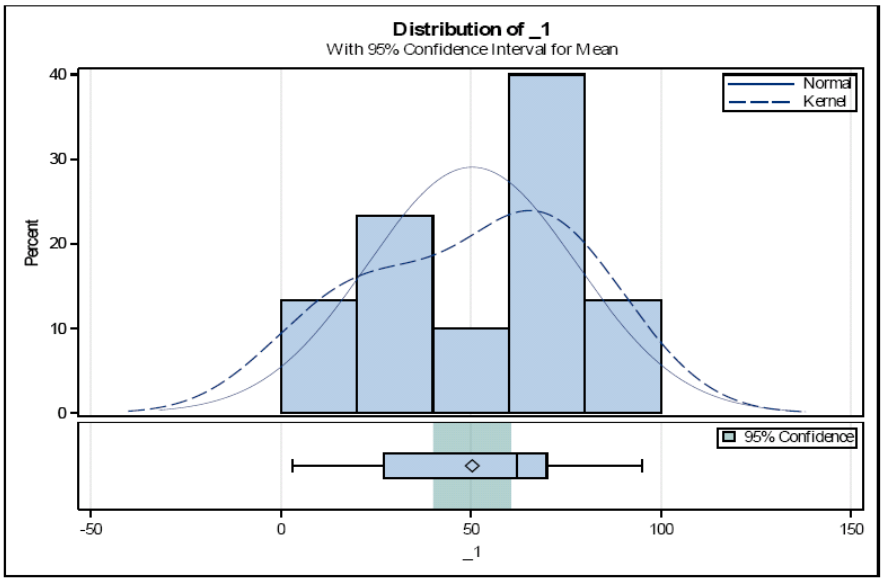


Figure 8 Q-Q Plot for the mean of the input

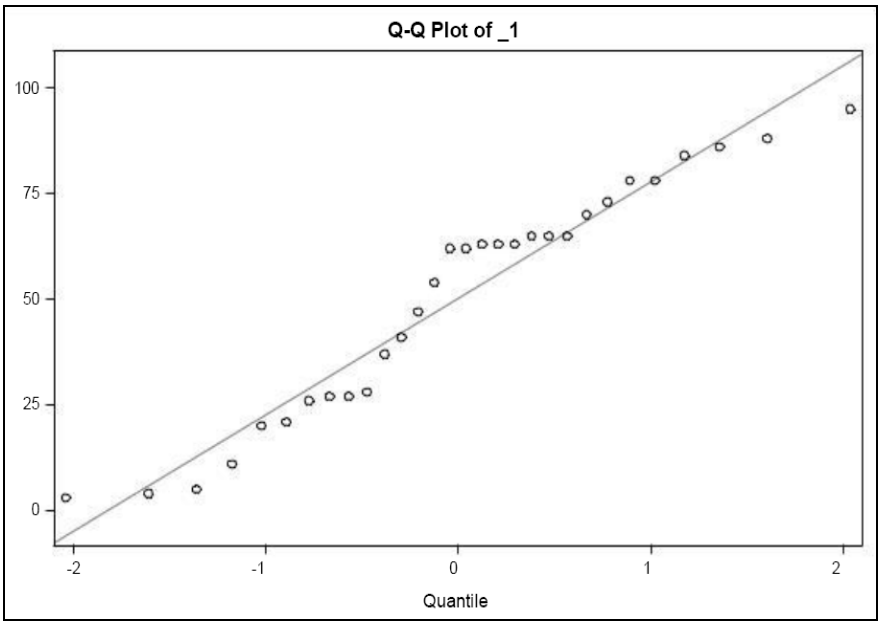


Figure 8 shows the Q-Q plot graph that shows the scattered nature of the words in the input. The T-test is done on one sample. This can be extended to all the samples in the dataset.

4.4.3 Correlation analysis

The correlation coefficient is a measure of linear association between the words present in two or more inputs or samples from the dataset. Values of the correlation coefficient are always between –1 and +1.

Table 5 shows the list of samples or inputs used for the analysis.

Table 5 Variables list for analysis

Two variables	_1	_2

Table 6 shows the analytical values obtained from the study of input variables _1 and _2.

Table 6 Simple statistics based on the two samples

Simple statistics						
Variable	*N*	*Mean*	*Std. dev.*	*Sum*	*Minimum*	*Maximum*
_1	30	50.36667	27.47348	1511	3.00000	95.00000
_2	29	47.55172	28.62764	1379	2.00000	92.00000

The correlation among the words present in the samples is examined using a statistical procedure Pearson correlation method and the correlation coefficients are tabulated in Table 7.

Table 7 Pearson correlation coefficients

Pearson correlation coefficients		
Prob. > |r| under H0: rho = 0		
Number of observations		
	_1	*_2*
_1	1.00000	–0.01375
		0.9436
	30	29
_2	–0.01375	1.00000
	0.9436	
	29	29

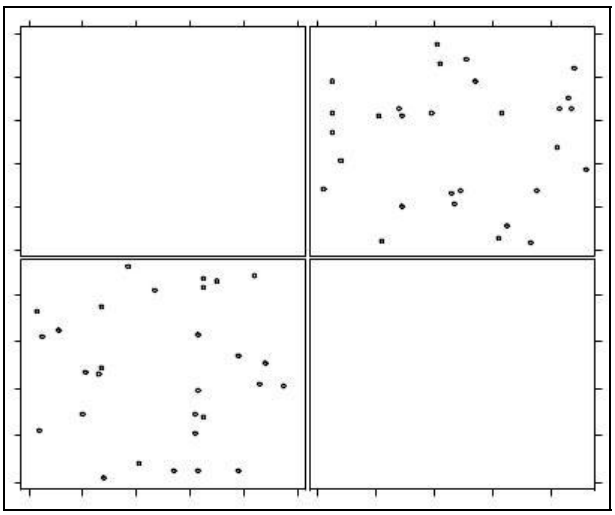
Figure 9 demonstrates the correlation or dependency of the inputs in the form of a scatter plot matrix. This test runs over the complete dataset and the results are shown in the form of a five-scale scatter plot matrix.

4.4.4 Bland-Altman analysis

The Bland-Altman analysis is a process to verify the extent of agreement or disagreement between two methods designed to measure the same parameters. A high correlation

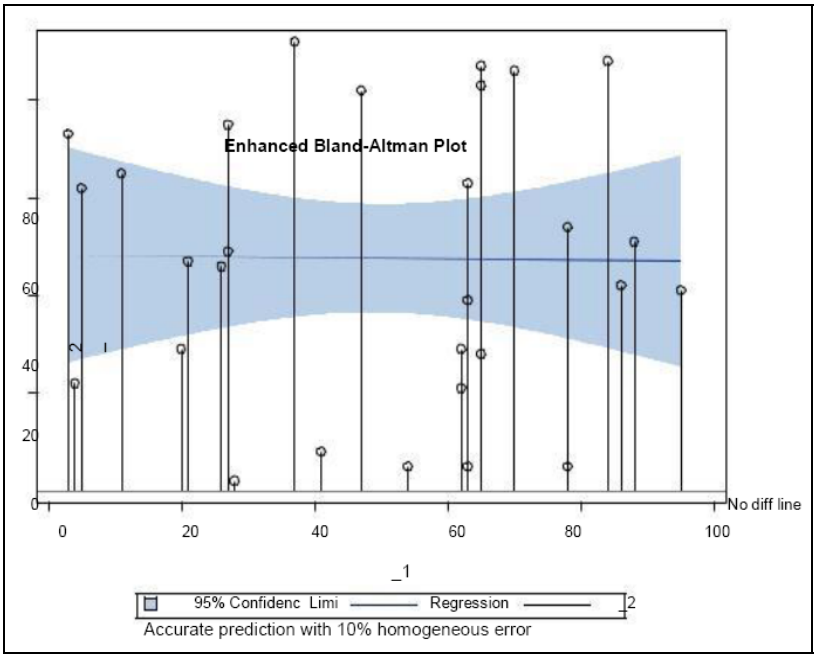
between the methods indicates that a better sample inputs has been chosen in data analysis.

Figure 9 Scatter plot matrix based on correlation values



The inputs given to the test are two samples. The graph shows a 10% homogenous error, i.e., a good level of interdependency between the samples is seen in Figure 10. An advanced analysis on the complete dataset is handled by selecting the structured Bland-Altman analysis procedure.

Figure 10 Enhanced Bland-Altman plot (see online version for colours)



4.4.5 Chi-square test

This test is used to examine the association between two categorical class labels. It can be used to test both extent of dependence and extent of independence between the words present in the inputs.

Table 8 shows the frequency of each word in the sample. The cumulative frequency and the cumulative percent are the values obtained with respect to the frequency and the percent values of the previous inputs in the sample. The test is extended to the dataset that produces the results based on the probability values.

Table 8 Chi-square frequency and percent values

_1	*Frequency*	*Percent*	*Cumulative frequency*	*Cumulative percent*
1	1	11.11	1	11.11
23	1	11.11	2	22.22
44	1	11.11	3	33.33
46	1	11.11	4	44.44
63	1	11.11	5	55.56
72	1	11.11	6	66.67
89	1	11.11	7	77.78
91	2	22.22	9	100.00
Frequency missing = 5				

4.4.6 Fisher's exact tests

Fisher's exact test is a statistical test used to determine if there are non-random associations between two categorical class labels. The high associations among the words in the input show that the data are highly interdependent and strongly associated. The analysis is done using two samples from the dataset. The test characteristics of the test are extended to the dataset using the scaling feature of the analysis procedure.

Table 9 Frequency count for two different samples

Table of _1 by _2															
_1	*_2*														
Frequency	*8*	*12*	*47*	*52*	*57*	*60*	*61*	*62*	*64*	*68*	*74*	*81*	*83*	*93*	*Total*
1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
23	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
44	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
46	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
63	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
72	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
89	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
91	0	0	0	0	1	0	0	1	0	0	0	0	0	0	2
Total	1	0	0	0	1	1	0	1	1	1	1	1	0	1	9
Frequency missing = 5															

Table 10 shows the varied nature of the data present the input based on the likelihood ratio and the phi coefficient values obtained as the results of the test. The probability value 1.0000 shows that the data possess more associations.

Table 10 Chi square likelihood values based on probability

Statistic	*DF*	*Value*	*Prob.*
Chi-square	56	63.0000	0.2425
Likelihood ratio chi-square	56	36.7775	0.9780
Mantel-Haenszel chi-square	1	0.4010	0.5266
Phi coefficient		2.6458	
Contingency coefficient		0.9354	
Cramer's V		1.0000	

Table 11 Fisher's exact test p-values

Fisher's exact test	
Table probability (P)	< .0001
Pr <= P	1.0000

This test uses chi-square test results to obtain the non-random associations between the words in the inputs. The table probability value shows that the sample _2 is the sample with more non-random associations.

4.4.7 One-way ANOVA

ANOVA stands for analysis of variance. It performs analysis of the data from a wide variety of experimental designs. In this process, a continuous response class label, known as dependent inputs, is measured under experimental conditions identified by classification class labels, known as independent inputs. The variation in the response is assumed to be due to effects in the classification, with random error accounting for the remaining variation.

Table 12 shows the list of distinct values present in the input variable _1 and the distinct count is stated as the number of levels.

Table 12 Class level information of ANOVA

Class level information		
Class	*Levels*	*Values*
_1	23	3 4 5 11 20 21 26 27 28 37 41 47 54 62 63 65 70 73 78 84 86 88 95

The analytical results of the count are shown in Table 13.

Table 13 Number of observations analysis using ANOVA

Number of observations read	35
Number of observations used	29

The DF and probability values are presented in Table 14 that shows the probability of the error.

Table 14 Sum of square analysis using ANOVA

Source	*DF*	*Sum of squares*	*Mean square*	*F-value*	*Pr > F*
Model	21	17,504.00575	833.52408	1.07	0.4978
Error	7	5,443.16667	777.59524		
Corrected total	28	2,2947.17241			

The analysis method arrived at an error value and mean value based on the variable condition given and the results are shown in Table 15.

Table 15 R-square analysis using ANOVA

R-square	*Coeff. var.*	*Root MSE*	*_2 mean*
0.762796	58.64224	27.88539	47.55172

The ANOVA between the two variables or inputs _1 and _2 is represented in the terms of DF and F values.

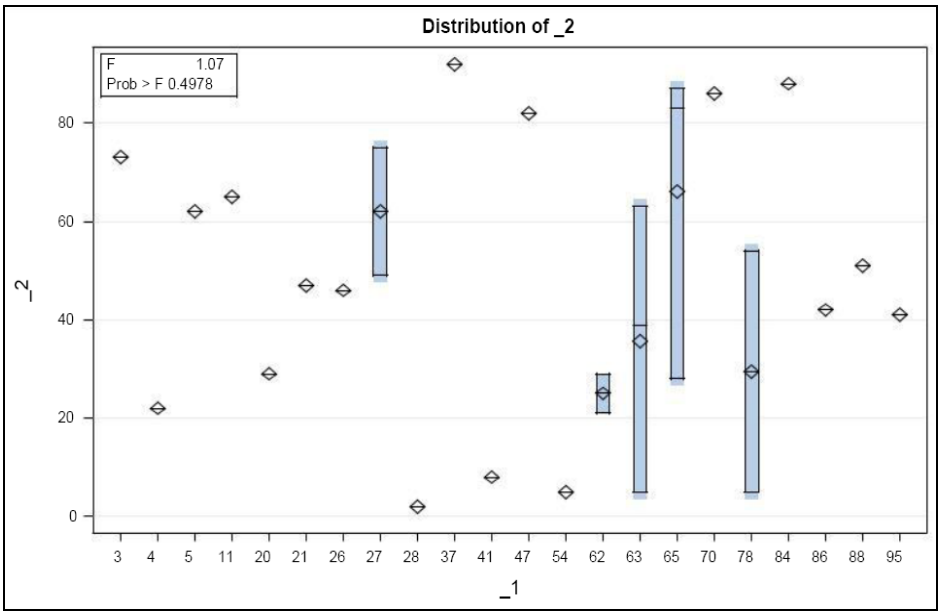
The results are shown in Table 16.

Table 16 Results of ANOVA DF

Source	*DF*	*ANOVA SS*	*Mean square*	*F-value*	*Pr > F*
_1	21	17504.00575	833.52408	1.07	0.4978

Figure 11 shows the box plot graph input sample _1 with respect to the input sample _2. The bars represent the range of values. The diagonals and the point lines show the mean and median values of the inputs. The F value and the probability values are presented in the graph. This statistics is extended the dataset using the multi way ANOVA and enhanced box plot graph.

Figure 11 Bar chart result of ANOVA (see online version for colours)



The statistical analysis is carried on the input data to understand the various features of data like scatter, variance, interdependency and deviation. The results show that the data is highly interdependent and correlated.

5 Conclusions and future work

The proposed system works on detection of cyberbully in crime investigation forums. In an era of internet and social networking, a hectic task of cyberbully detection is made easy by this framework. The use of intelligent algorithms like MCC, the size of the dataset is greatly reduced which in turn reduces the computational time and complexity. The cyberbully words identified are classified into five classes as vulgar, bad, terrorism, threatening and insulting using SVM. The system is made dynamic by automatically deducting the posts from the forum. The results drawn by the system are generated as a report so that is useful for taking the further action on the cyberbully. This proposed framework shows better results while the action is to stop the online users becoming the victims of cyberbully.

The scope of the paper can be further increased by encrypting the results obtained from the system. A vital and non-ethical problem that the technology is facing is 'hacking'. So, the role of Information Security has become much more important. To ensure the reliability of the results obtained from the system, they are stored in an encrypted format. But since the system is dynamic, the encryption algorithm should also be dynamic in nature. All these advance features make the system sophisticated and maintainable.

References

Abdelhaq, H., Gertz, M. and Armiti, A. (2016) 'Efficient online extraction of keywords for localized events in twitter', *GeoInformatica*, Vol. 21, No. 2, pp.365–388.

Al-garadi, M.A., Varathan, K.D. and Ravana, S.D. (2016) 'Cybercrime detection in online communications: the experimental case of cyberbullying detection in the Twitter network', *Computers in Human Behavior*, Vol. 63, pp.433–443.

Bilski, P. (2014) 'Data set preprocessing methods for the artificial intelligence-based diagnostic module', *Measurement*, Vol. 54, pp.180–190.

Chen, Y., Zhou, Y., Zhu, S. and Xu, H. (2012) 'Detecting offensive language in social media to protect adolescent online safety', in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, September, pp.71–80, IEEE.

Dadvar, M., de Jong, F.M., Ordelman, R.J.F. and Trieschnigg, R.B. (2012) 'Improved cyberbullying detection using gender information', in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, University of Ghent.

Kowalski, R.M., Giumetti, G.W., Schroeder, A.N. and Lattanner, M.R. (2014) 'Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth', *Psychological Bulletin*, July, Vol. 140, No. 4, p.1073.

Kowalski, R.M., Limber, S.P. and Agatston, P.W. (2009) 'Cyberbullying: bullying in the digital age', *Journal of Blackwell Publishing*, July, Vol. 7, No. 2, pp.9–17.

Lossio-Ventura, J.A., Jonquet, C., Roche, M. and Teisseire, M. (2016) 'Biomedical term extraction: overview and a new methodology', *Information Retrieval Journal*, Vol. 19, Nos. 1–2, pp.59–99.

McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A. and Jakubowski, E. (2011) 'Learning to identify internet sexual predation', *International Journal of Electronic Commerce*, Vol. 15, No. 3, pp.103–122.

Orencik, C., Selcuk, A., Savas, E. and Kantarcioglu, M. (2016) 'Multi-keyword search over encrypted data with scoring and search pattern obfuscation', *International Journal of Information Security*, Vol. 15, No. 3, pp.251–269.

Pradheep, T., Yogeshwaran, T., Sheeba, J.I. and Devaneyan, S.P. (2018) 'Impulsive intermodal cyber bullying recognition from public nets', *International Journal of Advanced Research in Computer Science*, Vol. 9, No. 3, pp.59–63.

Prathyusha, T., Hemavathy, R. and Sheeba, J.I. (2017) 'Cyberbully detection using hybrid techniques', in *International Conference on Telecommunication, Power Analysis and Computing Techniques (ICTPACT 2017)*, pp.1–6, IEEE.

Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q. and Mishra, S. (2018) 'Scalable and timely detection of cyberbullying in online social networks', *SAC 2018, Symposium on Applied Computing*, ACM, pp.1738–1747, ISBN: 978-1-4503-5191-1/18/04.

Raisi, E. and Huang, B. (2018) 'Weakly supervised cyberbullying detection using co-trained ensembles of embedding models', in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, August, pp.479–486, IEEE.

Sheeba, J.I. and Devaneyan, S.P. (2016) 'Cyberbully detection using intelligent techniques', *International Journal of Data Mining and Emerging Technologies*, Vol. 6, No. 2, pp.86–94.

Sheeba, J.I. and Devaneyan, S.P. (2018) 'Cyberbullying among adolescents: an effort of identification, prevention and intervention of cyberbullying', *JIMS8I – International Journal of Information Communication and Computing Technology*, Vol. 6, No. 1, pp.318–324.

Ybarra, M. (2010) 'Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression', *National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda*.

Zhao, R., Zhou, A. and Mao, K. (2016) 'Automatic detection of cyberbullying on social networks based on bullying features', in *Proceedings of the 17th International Conference on Distributed Computing and Networking*, January, p.43, ACM.