Chapter

Classification Model for Bullying Posts Detection

K. Nalini and L. Jabasheela

Abstract

Nowadays, many research tasks are concentrating on Social Media for Analyzing Sentiments and Opinions, Political Issues, Marketing Strategies and many more. Several text mining structures have been designed for different applications. Harassing is a category of claiming social turmoil in different structures and conduct toward a singular or group, to damage others. Investigation outcomes demonstrated that 7 young people out of 10 become the casualty of cyber bullying. Throughout the world, many prominent cases are existing due to the bad communications over the Web. So there could be suitable solutions for this problem and there is a need to eradicate the lacking in existing strategies in dealing problems with cyber bullying incidents. A prominent aim is to design a scheme to alert the people those who are using social networks and also to prevent them from bullying environments. Tweet corpus carries the messages in the text as well as it has ID, time, and so forth. The messages are imparted in informal form and furthermore, there is variety in the dialect. So, there is a requirement to operate a progression of filtration to handle the raw tweets before feature extraction and frequency extraction. The idea is to regard each tweet as a limited blend over a basic arrangement of topics, each of which is described by dissemination over words, and after that analyze tweets through such topic dispersions. Naturally, bullying topics might be related to higher probabilities for bullying words. An arrangement of training tweets with both bullying and non-bullying texts are required to take in a model that can derive topic distributions from tweets. Topic modeling is used to get lexical collocation designs in the irreverent content and create significant topics for a model.

Keywords: cyberbullying, Twitter, LDA, SVM, TF-IDF

1. Introduction

The proposed methodology is a dual compound method. It utilizes the arrangement of "bullying" or "non-bullying" class and also it utilizes link analysis to locate the most dynamic users as predators and victims. Each step can be explained in detail as follows. The feature selection is an essential phase in denoting data within component space to the categorizers. Mostly the data available from social network are noisy. So, there is a need to apply pre-processing techniques in order to obtain the research data with better quality followed by successive systematic steps; Moreover, sparsity in feature space increases with the count of documents. Nevertheless, the following types of features generated through the B-LDA topic model and weighted B-TF-IDF scheme. In the initial step, semantic highlights are related for locating harassing, abusive and offending posts. In pestering discovery the presence of pronouns in the nuisance post was represented. Essentially in this work, three sorts of capabilities are utilized. They are depicted as follows: (i) all second individual pronouns "you," "yourself," and so forth are considered one term; (ii) all other outstanding pronouns "he," "she," and so on., are viewed together as another element; (iii) foul words such as "fr**k," "shit," "moronic," and so forth., which make the post merciless are assembled in another arrangement of highlights. The new harassing words lexicon was made in view of the accompanying essential sites like *noswearing.com and urban dictionary*. The primary rationale behind consolidating these features is that it will boost the viability of the classification of tormenting posts. The classification outcomes are revealed in the experiments.

2. Review of literature

Rahat et al. [1] presented a multi-stage cyber bullying detection results that radically decreases the classification period and give warning signals. The system is greatly scalable without forfeiting precision and highly approachable in raising signals. It also contained an active priority scheduler and a rising classification procedure by applying Vine data sets. The performance outcomes demonstrate that the model enhances the scalability of digital harassing discovery contrasted to non-priority model and also explained that the system could fully check Vine-scale networks. The results depict that this digital harassing detection is considerably more measurable and receptive than the present modern technology. Zhong et al. [2] proposed an investigation to find out cyberbullying in Instagram utilizing the improvement of early-warning methods to detect offensive images.

The research operated by obtaining a huge volume of pictures in the Instagram image sharing process along with messages. They studied new features of the topics acquired from the picture portrayal and trained using neural network technology, added with images and texts. The results got the potential objectives for harassing on the characterization of texts and images. Sherly [3] proposed research using supervised feature selection to select the characteristics from the tweets by the ranking method. Then extreme learning machine (ELM) classifier is applied to execute the cyberbullying detection and enhance the precision and reduce the performance period. The performance investigation of the SFS-ELM model observed that the accuracy is improved by 13% and executed using MATLAB. Micheline et al. [4] accomplished a study by using an unsupervised methodology to identify harassing messages in social networks, utilizing Growing Hierarchical Self Organizing Map. The research contains various features to find semantic and syntactic interactions of regular cyber tormentors. They conducted various trials on FormSpring, Twitter and YouTube networks by collecting real time datasets. The outcomes of the research show that the model attains the significant performance and also promotes permanent watching applications to alleviate the huge issues of harassing. Suchini et al. [5] applied a text classification model to categorize the text as insulting or not. Feature selection is performed using Chi-square test and then classification algorithms are utilized for segregating comments as insulting or non-insulting words. Various algorithms like SVM, Naive Bayes, Logistic Regression, Random Forest are applied and out of all algorithms, SVM gave better results.

Krishna et al. [6] proposed a model deployed for detecting abusive text and images in the social network. This automated system could find the offensive content in messages using the combination of a bag of visual word method, local binary pattern and SVM classifier. The offensive detection in the text messages are executed by a bag of word method with Naïve Bayes classifier and then the Boolean system is applied to classify the content. Javier et al. [7] have displayed automatic strategies for identifying erotic plundering in Chat rooms. They have effectively demonstrated that a learning-based technique is an attainable method to approach this issue and have proposed novel sets of highlights to determine the classification of chat partakers as exploiters or non-exploiters. They exhibited that the arrangements of features used and the comparative weighting of the disarrangement expenditures in the SVMs are two fundamental factors that ought to be considered to upgrade execution.

Huang et al. [8] proposed normal text investigation using social network characteristics to classify harassing in Twitter and also considered the social connection between clients would betterment outcome for classification. Zhao et al. [9] applied a collection of features known as EBoW (Natural Language Processing method), containing a bag of words structure connected with Latent Semantic analysis and word embeddings by computing word vectors. They also used SVM to classify the data collection in Twitter which contains keywords like bully or bullying.

Chen et al. [10] researched existing content mining techniques in recognizing harassing texts for ensuring adolescent online safety. In particular, they proposed the Lexical Syntactical Feature (LSF) way to deal with hostile contents on the internet and further foresee a client's potentiality to convey hostile contents. Their investigation has many commitments. To begin with, they essentially conceptualize the idea of online hostile contents and further recognize the contribution of pejoratives/obscenities and profanities in deciding offensive substance, and present hand creating syntactic standards in finding verbally abusing provocation. Second, they enhanced customary Machine-Learning strategies by not just utilizing lexical features to identify hostile dialect, yet in addition style feature, structure features, and content-specific features to better foresee a client's possibility to convey hostile content in social media. Investigation result demonstrates that the LSF Sentence offensiveness forecast and client offensiveness estimate algorithm beat, customary learning-based methodologies in turns of precision, recall, and F-score. The LSF endures casual and incorrect spelling contents and it can possibly adjust to any forms of English written word styles.

3. The Bully-latent Dirichlet allocation (B-LDA): model design

LDA is an outstanding method of Bayesian multinomial mixture model in text analysis based on its ability to assemble, elucidate and semantically cogent topics. It uses the Dirichlet distribution to model the distribution of the topics for each and every one document. In LDA, each word is measured from a multinomial distribution over words particular to this topic. Since LDA is extremely modular and hierarchical, consequently, it can simply be broadened. Various expansions to basic LDA model have been recommended to incorporate document metadata. The easy process of integrating the metadata in generative topic models is to create both the words and the metadata concurrently specified unseen topic variables. The Author-Topic (AT) model resembles Bayesian network, in which every authors' attractions are modeled with a combination of topics [11]. In this model an arrangement of authors, advertisements are watched and looked over from different documents depends on their topics. To create each word, an author x is picked at identical from this set, then a topic z is chosen from a topic distribution θ_x that is particular to the author, and after that, a word w is created by testing from a topic-particular multinomial distribution ϕ_z.

The proposed Bully-LDA (B-LDA) model is used for identifying bullying words used by authors. This model captures bullying-topics which are used in social networks like Twitter. In Twitter, one person sends tweets to many followers. Here in this model, the sender is considered as Predator, when he/she sends bullying words to their followers. The followers are represented as Victims. The B-LDA model is a generative process model and also encapsulates topics and the communication networks of Predators and Victims by conditioning the multinomial distribution over bullying topics distinctly on both the Predator and a Victim of a bullying message. Unlike other models, B-LDA model takes into concern both predator and victims distinctly. The motive of the predator is also considered in addition to this representation. Each motive is associated with a set of topics, and these topics may overlap. For example, the categories of motive can be racist, sexual, outrage, irrelevant. The sexual motive of predator contains the topics of crude, implicit/ambiguous language or an indecent proposal. The Racist category contains more abusive matters such as homophobia, extremism, slurs, etc. The outrage is a category, which specifies reactions that express contempt. The messages that do not contain any form of offensive language are considered to be irrelevant. Each predator has a multinomial distribution over motives. Thus, B-LDA model is a clustering model, in which appearances of topics are the underlying data, and sets of correlated topics are together gathered as clusters that denote motive. Predators and Victims are mapped to motive assignments, and then a topic is selected based on these motives. The intention of each and every predator has a multinomial distribution on topics, and every topic has a multinomial distribution on words. First, the motive assignments can be made separately for each word in a document. This model represents that someone can change motive during the exchange of the messages.

Author-Topic (AT) [11] model has been extended by incorporating a new set of variables like authors as Predators and Victims, the motivation of an author. In this generative process for each message, a Predator, p_d and a set of Victims, v_d are observed. To generate each word, a victim y is chosen at uniform from v_d, and then a motive x for the Predator is chosen from multinomial motive distribution ψpd. Next a topic z is selected from a multinomial topic distribution θ_x, in which the

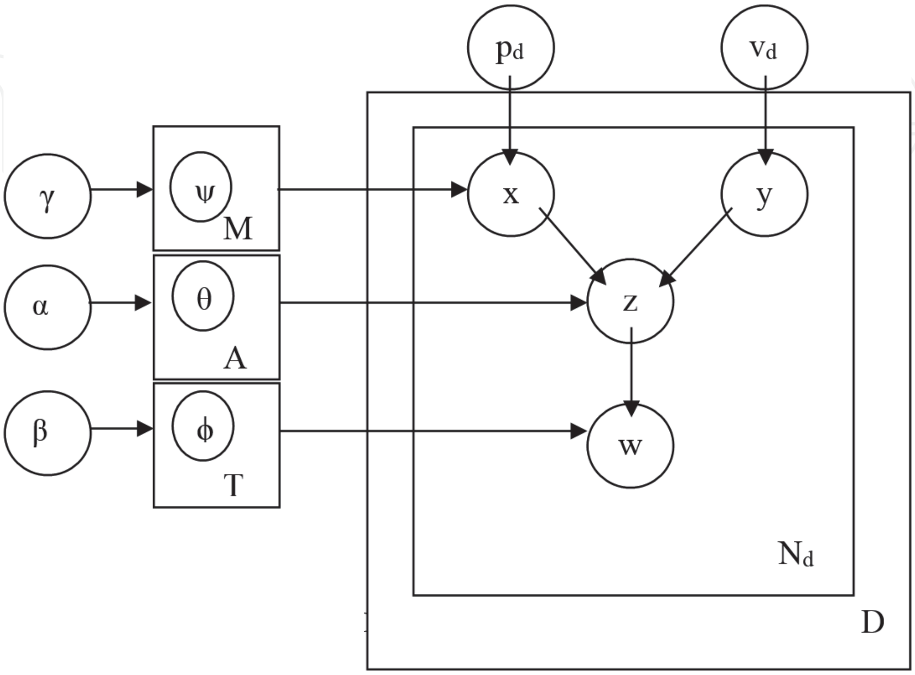


Figure 1.
Graphical model for B-LDA.

distribution is specific to the predator-motive(x). At last, the word w is produced by sampling from a topic-meticulous multinomial distribution ϕ_z.

Figure 1 is a schematic diagram of the B-LDA model.

The generative procedure of this strategy is as follows:

1. for every motive m with m = 1, … ..M, choose $\psi_m \sim \mathrm{Dir}(\gamma)$

2. for each predator and victim pair (x,y) with x = 1, … .,A and y = 1, … .,A choose $\theta x,y \sim$ Dir (α)

3. for each topic t with t = 1, … ..T, choose $\phi_t \sim \mathrm{Dir}(\beta)$

4. for each message d

   a. observe motive m_d

   b. observe predator p_d and the victims v_d

   c. for each word w in d

      i. choose topic $z_{dn} \sim \theta_{zd}$

      ii. choose word $w_{dn} \sim \phi_{zdn}$

In this model for a particular message d, given the hyper parameters α, β, and γ, the predator p_d, and set of victims v_d, the connected dispersion of an author blend θ, a motive blend ψ, a topic blend ϕ, a set of N_d victims y_d, and a set of N_d predator motives x_d, a set of N_d topics z_d and a set of N_d words w_d is assigned by,

$$\begin{aligned} p(\theta,\phi,\psi,yd,xd,zd,wd|\alpha,\beta,\gamma,pd,vd) \\ &= p(\psi|\gamma)p(\theta|\alpha)p(\phi|\beta) \\ &= \prod_{n=1}^{Nd} p(ydn|vd)p(xdn|pd)p(zdn|\theta xdn)p(wdn|\phi zdn) \end{aligned} \tag{1}$$

Integrating over γ, θ and ϕ and summing over y_d, x_d, and z_d, the marginal distribution of a document is calculated as follows:

$$\begin{aligned} p(wd|\alpha,\beta,\gamma,pd,vd) = \iiint p(\psi|\gamma)p(\theta|\alpha)p(\phi|\beta) \\ \prod_{n=1}^{Nd}\sum_{ydn}\sum_{xdn}\sum_{zdn} p(ydn|vd)p(xdn|pd)p(zdn|\theta xdn)p(wdn|\phi zdn)d\psi d\phi d\theta \end{aligned} \tag{2}$$

Then the product of the marginal probabilities of single documents, and the probability of a corpus is computed as,

$$p(D|\alpha,\beta,\gamma,p,v) = \prod_{d=1}^{D} p(wd|\alpha,\beta,\gamma,pd,vd) \tag{3}$$

3.1 Monte Carlo Gibbs sampling

The assumption on models in the LDA family cannot be carried out correctly. Three standard approximations have been occupied to acquire practical results:

Variational methods [12], Gibbs sampling [13], and expectation propagation [14]. As Gibbs sampling is easy to implement, it has been applied here. There is a need to derive a formula to carry out the Gibbs sampling for $P(z_i,y_i,x_i|z_{-i},y_{-i},x_{-i})$, the conditional distribution of a topic and victims for w word given all other words topic and victim assignment, the motive of the predator, z-i, y-i, and x-i. In order to calculate P(z,y,x|w), the posterior distribution of topic, victim assignments and the motive of the predator given the words in the corpus.

The calculations begin with P(w|z,x), using P(w|z,x,Φ) in order to integrate out the unknown Φ distributions to obtain: $P(w|z,y,\Phi) = \prod_{iw=1}^{W} \phi ziw(Wiw)$.

Reorganizing the product over the W word token exist in the corpus to collect words that are assigned to the same bullying topic,

$$P(w|z,y,\Phi) = \prod_{z=1}^{T}\prod_{u=1}^{U}\phi_z^{n_z^{wu}} \tag{4}$$

where $n_z{}^{wu}$ is the number of times that a bullying word, w_u was assigned to a bullying topic. To integrate out the ϕ distribution by using the Dirichlet distributions,

$$\begin{aligned} p(w|z,y) &= \int \prod_{z=1}^{T}\left(\frac{\Gamma\left(\sum_{u=1}^{U}\beta u\right)}{\prod_{u=1}^{U}\Gamma(\beta u)}\left(\prod_{u=1}^{U}\phi_z^{n_z^{wu}+\beta u-1}(wu)d\phi_z(wu)\right)\right) \\ &= \prod_{z=1}^{T}\left(\frac{\Gamma\left(\sum_{u=1}^{U}\beta u\right)}{\prod_{u=1}^{U}\Gamma(\beta u)}\left(\frac{\prod_{u=1}^{U}\Gamma\left(n_z^{wu}+\beta u\right)}{\Gamma\left(\sum_{u=1}^{U}\beta u+\sum_{u=1}^{U}n_z^{wu}\right)}\right)\right) \end{aligned} \tag{5}$$

In the same manner, P(z,y) is computed using a procedure analogous to that used for P(w|z,y). The collected terms of bullying words are assigned to the same topic and predator-victim pair and integrate out the Θ distributions corresponding to all the different predator-victim pairs, P:

$$P(z,y) = \left(\prod_{iw=1}^{W}\frac{1}{n_R(diw)}\right)\prod_{p=1}^{P}\left(\frac{\Gamma\left(\sum_z \alpha z\right)}{\prod_{z=1}^{T}\Gamma(\alpha z)}\frac{\prod_z\Gamma\left(n_p^z+\alpha_z\right)}{\Gamma\left(\sum_z\alpha_z+\sum_z n_p^z\right)}\right) \tag{6}$$

where $nR(diw)$ is the number of victims corresponding to a word in a message.

Similarly can calculate P(z, x) using a procedure analogous to that used for P (w|z, x). Bullying words have been assigned to the same topic and the motivation of the predator can be computed as,

$$P(z,x) = \left(\prod_{iw=1}^{W}\frac{1}{n_S(diw)}\right)\prod_{p=1}^{P}\left(\frac{\Gamma\left(\sum_z \gamma_z\right)}{\prod_{z=1}^{T}\Gamma(\gamma z)}\frac{\prod_z\Gamma\left(n_m^z+\gamma_z\right)}{\Gamma\left(\sum_z\gamma_z+\sum_z n_m^z\right)}\right) \tag{7}$$

where $nS(diw)$ is the number of predators having bad motivation with respect to the bullying word in a message. An expression for P (w, z, y, x) can be achieved by combining the equations of P(w|z, y), P(z, y) and P(z, x). This can be used to write an expression for the posterior distribution of z, y and x given the corpus,

$$P(z,y,x|w) = \frac{P(w,z,y,x)}{\sum_{z,y,x}P(w,z,y,x)} \tag{8}$$

Hence the denominator cannot be calculated directly. The following equations are used to run a MCMC Gibbs sampling calculation by using the conditional distribution P(z$_i$, y$_i$, x$_i$, w$_i$|z$_{-i}$, y$_{-i}$, x$_{-i}$, w$_{-i}$).

$$
\begin{aligned}
&P(zi, yi, xi, wi|z-i, y-i, x-i, w-i)\\
&\quad= \frac{P(z,y,x,w)}{P(z-i,y-i,x-i,w-i)}\\
&\quad= \frac{1}{nR}\left(\frac{\dfrac{\Gamma\left(n_m^t+\gamma t\right)}{\Gamma\left(\sum_z n_m^z+\sum_z \gamma z\right)}}{\dfrac{\Gamma\left(n_m^t-1+\gamma t\right)}{\Gamma\left(\sum_z n_m^z-1+\sum_z \gamma z\right)}}\,\frac{\dfrac{\Gamma\left(n_p^t+\alpha t\right)}{\Gamma\left(\sum_z n_p^z+\sum_z \alpha z\right)}\dfrac{\Gamma\left(n_t^{wu}+\beta u\right)}{\Gamma\left(\sum_u n_t^{wu}+\sum_u \beta u\right)}}{\dfrac{\Gamma\left(n_p^t-1+\alpha t\right)}{\Gamma\left(\sum_z n_p^z-1+\sum_z \alpha z\right)}\dfrac{\Gamma\left(n_t^{wu}-1+\beta u\right)}{\Gamma\left(\sum_u n_t^{wu}-1+\sum_u \beta u\right)}}\right)\\
&\quad= \frac{1}{nR}\frac{n_{m,-i}^t+\gamma t}{\sum_z n_{m,-i}^z+\sum_z \gamma z}\frac{n_{p,-i}^t+\alpha t}{\sum_z n_{p,-i}^z+\sum_z \alpha z}\frac{n_{t,-i}^{wu}+\beta u}{\sum_u n_{t,-i}+\sum_u \beta u}
\end{aligned}
\tag{9}
$$

where the victim, y is part of Predator-Victim pair, *p*, the –i subscript is used to denote that the counts are taken by excluding the assignment of word *i* itself, and n_R is the number of Victims for the message to which word *i* belongs.

3.2 Experiments and results

In this chapter, the experimental results are discussed. The datasets used in these experiments are tweets from Twitter. An experiment has been conducted on tweets based on the architecture of an automatic cyber bullying detection system. Search is made in the Twitter stream for Tweets containing the strings that contain offensive words so as to particularly filter for tweets related to bullying. In total, more than 1,00,000 tweets are gathered between Jan 1st, 2015 and Jan 30th, 2016. A limit number of tweets are matching with the query. So, approximately 300 tweets are filtered per day. The statistics for training and the testing corpus is given in **Table 1**. Tweets were manually labeled as belonging to one of the different motives namely Sexual, Racist, Outrage, Irrelevant, and Unknown after the preprocessing. The examples of harassing comments posted on Twitter are listed below and depicted in **Figure 2(a)** and **(b)** and top bullying words which are extracted are given in **Table 2** (**Figures 3**–**5**).

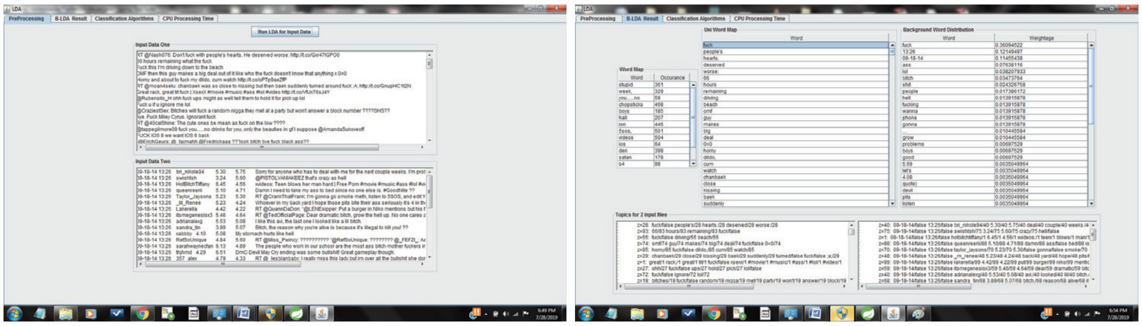
Date	Time	Tweets
01–13-15	12:16	NefarioussNess Do not fuck with people's hearts
09–18-15	11:51	TittyCityClay it's always been a self respect thing. Shit like this is stupid as fuck lol
05–13-15	10:11	djkeneechi Nah kiss no one ass to stay in my life anymore im tired of that shit it's time for me to man up

3.3 Results and discussions

Bully-Latent Dirichlet Allocation model is an intended for pictorial representation of texts in a harassing message, given their predator and a pair of casualties.

	Training corpus	Testing corpus
Tweets	3,18,14,716	97,35,537
Retweets	76,20,335	2,87,567
URLs	85,45,112	4,76,234
Usernames	97,02,445	14,20,554
Hashtags	79,85,956	3,56,778

Table 1.
Statistics of training and testing corpus.



Input Data One

```
RT @Nash076: Don't fuck with people's hearts. He deserved worse: http://t.co/Gxr47IGPO0
66 hours remaining what the fuck
Fuck this I'm driving down to the beach
OMF then this guy makes a big deal out of it like who the fuck doesn't know that anything x 0=0
Horny and about to fuck my dildo, cum watch http://t.co/oPTp9aeZfP
RT @moan4selu: chanbaek was so close to kissing but then baek suddenly turned around fuck ;A; http://t.co/GnupHC1t2N
Great rack, great tit fuck | XsexX #movie #music #ass #lol #video http://t.co/VfUx76sJ4Y
@Rubensito_H ohh fuck ups might as well tell them to hold it for pick up lol
Fuck u if u ignore me lol
@CraziestSex: Bitches will fuck a random nigga they met at a party but won't answer a block number ????SHS??
Aye. Fuck Miley Cyrus. Ignorant fuck
RT @40calShine: The cute ones be mean as fuck on the low ????
@tappegilmore09 fuck you.....no drinks for you..only the beauties in gf I suppose @AmandaSukoveoff
FUCK IOS 8 we want IOS 6 back
@ErichGeurs: @_tajmahh @Fredrickaaa ??"look bitch bye fuck black ass??
```

Input Data Two

```
09-18-14 13:26  bri_nikole94      5.30  5.75  Sorry for anyone who has to deal with me for the next couple weeks. I'm prot
09-18-14 13:26  swishtsh          3.24  5.60  @PISTOLVANIAKEEZ that's crazy as hell
09-18-14 13:26  HotBitchTiffany   6.45  4.56  xvideos: Teen blows her man hard | Free Porn #movie #music #ass #lol #vi
09-18-14 13:26  queenreenii       5.10  4.71  Damn I need to take my ass to bed since no one else is. #GoodNite ??
09-18-14 13:26  Taylor_Jaysons    5.23  5.30  RT @CrankThatFrank: I'm gonna go smoke meth, listen to 5SOS, and edit Y
09-18-14 13:26  _M_Renee          5.23  4.24  Whoever in my back yard I hope those pits bite their ass seriously it's 4 in th
09-18-14 13:26  Lakerella         4.42  4.22  RT @QuannDaDon: "@LENEsipper: Put a burger in Niko mentions but his f
09-18-14 13:26  itsmegenesisx3    5.46  4.64  RT @TedOfficialPage: Dear dramatic bitch, grow the hell up. No one cares a
09-18-14 13:26  adrianalexg       5.53  5.08  I like this avi, the last one I looked like a lil bitch.
09-18-14 13:26  sandra_tin        3.89  5.07  Bitch, the reason why you're alive is because it's illegal to kill you! ??
09-18-14 13:26  xabbby   4.10     5.08  My stomach hurts like hell
09-18-14 13:26  RefSoUnique       4.84  5.60  RT @Miss_Pwincy: ?????????? "@RefSoUnique: ???????? @_FEFZI_: Az
09-18-14 13:26  sarahwaynecfan    6.13  4.89  The people who work in our school are the most ass bitch mother fuckers ir
09-18-14 13:26  thjhmd   4.29     6.51  DmC Devil May Cry ending was some bullshit! Great gameplay though.
09-18-14 13:26  357_alex          4.79  4.33  RT @_lesbianbaby: I really miss this lady,but im over all the bullshit she dor
```

(a)

Word Map

Word	Occurance
stupid	361
week,	329
you.....no	58
chopsticks	468
boys	185
hall	207
ion	446
5sos,	501
videos	504
ios	64
den	398
satan	178
b4	98

Uni Word Map

Word
fuck
people's
hearts.
deserved
worse:
66
hours
remaining
driving
beach
omf
guy
makes
big
deal
0=0
horny
dildo,
cum
watch
chanbaek
close
kissing
baek
suddenly

Background Word Distribution

Word	Weightage
fuck	0.36094522
13:26	0.12149497
09-18-14	0.11455438
ass	0.07638116
lol	0.038207933
bitch	0.03473764
shit	0.024326758
people	0.017386172
hell	0.013915878
fucking	0.013915878
wanna	0.013915878
phone	0.013915878
gonna	0.013915878
...	0.010445584
grow	0.010445584
problems	0.00697529
boys	0.00697529
good	0.00697529
5.69	0.0035049964
let's	0.0035049964
4.08	0.0035049964
quote)	0.0035049964
devil	0.0035049964
pits	0.0035049964
listen	0.0035049964

Topics for 2 input files

```
z=28:  fuck/false people's/28 hearts./28 deserved/28 worse:/28
z=93:  66/93 hours/93 remaining/93 fuck/false
z=55:  fuck/false driving/55 beach/55
z=74:  omf/74 guy/74 makes/74 big/74 deal/74 fuck/false 0=0/74
z=85:  horny/85 fuck/false dildo,/85 cum/85 watch/85
z=29:  chanbaek/29 close/29 kissing/29 baek/29 suddenly/29 turned/false fuck/false ;a;/29
z=1:  great/1 rack,/1 great/1 tit/1 fuck/false xsexx/1 #movie/1 #music/1 #ass/1 #lol/1 #video/1
z=27:  ohh/27 fuck/false ups/27 hold/27 pick/27 lol/false
z=72:  fuck/false ignore/72 lol/72
z=19:  bitches/19 fuck/false random/19 nigga/19 met/19 party/19 won't/19 answer/19 block/19
```

```
z=40:  09-18-14/false 13:26/false bri_nikole94/40 5.30/40 5.75/40 deal/40 couple/40 weeks./4
z=75:  09-18-14/false 13:26/false swishtsh/75 3.24/75 5.60/75 crazy/75 hell/false
z=1:  09-18-14/false 13:26/false hotbitchtiffany/1 6.45/1 4.56/1 xvideos:/1 teen/1 blows/1 man/1
z=88:  09-18-14/false 13:26/false queenreeni/88 5.10/88 4.71/88 damn/88 ass/false bed/88 is
z=70:  09-18-14/false 13:26/false taylor_jaysons/70 5.23/70 5.30/false gonna/false smoke/70
z=48:  09-18-14/false 13:26/false _m_renee/48 5.23/48 4.24/48 back/48 yard/48 hope/48 pits/
z=99:  09-18-14/false 13:26/false lakerella/99 4.42/99 4.22/99 put/99 burger/99 niko/99 mentic
z=59:  09-18-14/false 13:26/false itsmegenesisx3/59 5.46/59 4.64/59 dear/59 dramatic/59 bitc
z=40:  09-18-14/false 13:26/false adrianalexg/40 5.53/40 5.08/40 avi,/40 looked/40 lil/40 bitch./
z=68:  09-18-14/false 13:26/false sandra_tin/68 3.89/68 5.07/68 bitch,/68 reason/68 alive/68 il
```

(b)

Figure 2.
(a) Bullying words with their probability, and (b) List of bullying words.

Word	Prob	Word	Prob	Word	Prob
Fuck	0.0798	Bitch	0.0705	Naked	0.0588
Ass	0.0767	Freak	0.0699	Sexy	0.0569
shit	0.0752	Fat	0.0663	Mood	0.0547
Gay	0.0738	Dirty	0.0643	Lick	0.0519
Dumb	0.0722	Bullshit	0.0621	Bed	0.0508
Suck	0.0711	Kiss	0.0604	Piss	0.0495

Table 2.
Extracted top bullying words.

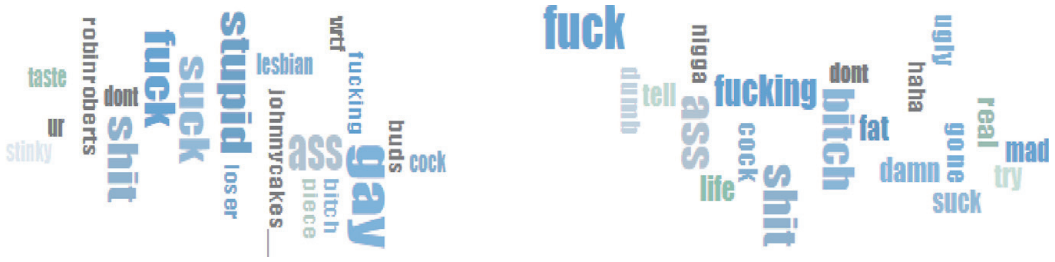


Figure 3.
Word cloud for bullying words.

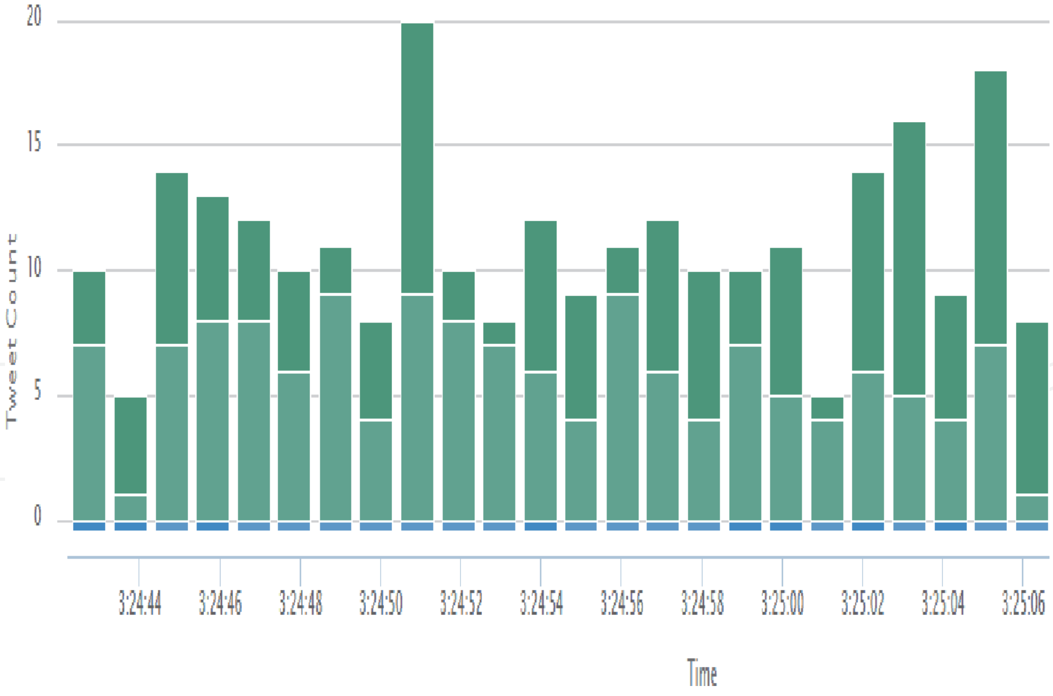


Figure 4.
Number of bullying tweets over time intervals.

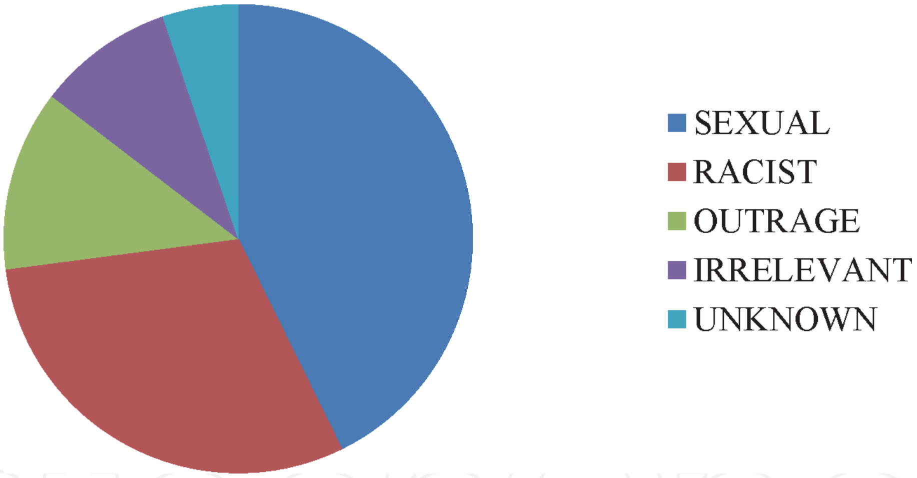


Figure 5.
Distributions of tweets per motive.

B-LDA got crucial enrichment to facilitate specification the per-bullying message topic dispersion mutually on the predator and individual victims. Every topic includes multinomial distribution on words and every Predator-Casualty pair has a distribution on topics. So, subsidiary dispersions in excess on bullying subjects accustomed exclusively on a predator, or solely on a recipient, can be computed easily. For example, corpus comprising 135 persons and 35 k bulling messages, and also on 5 months of sending and receiving messages of a predator, comprising 17 victims and 19 k messages. B-LDA turns up tremendously prominent topics, and grants support that it predicts predator's motives. In the experiments, the hectic parameters α and β are fixed at 1 and 0.01 respectively. The number of topics T is also fixed at |T| = 5. For a 50 topic solution, Dataset from Twitter took 150 hours for 2000 iterations (5 min per iteration).

B-LDA proves the motive of the predator and track the activity of the predator with victims, using the following steps. First, the proportions of each predator contributing in each of the bullying topics are determined. Next the impacts of the predators throughout the time intervals on the bullying topics. The two users' threshold ε and λ are empirically set to 3.2106 and 2.0457, respectively. From each of the documents, B-LDA generates 5 topics with predators associated with each. The distribution of the different bullying topics from the documents is displayed in **Table 3**. From the table, predator p1 has a probability of 0.0547 for bullying topic t5. There is a need to prove the bullying motive of the predator with victim using specific time intervals within bullying topics. It could be characterized as trails: A tweet message is a triplet (a, μ, τ), representing a textual bullying message μ written by the predator "a" at time τ. A document, denoted by d, is a sequence of bullying messages ordered by τ. From this definition, time τ_d is associated with both message μ_d and predator a_d.

The predator time contributions during time interval have been evaluated by:

$$F\left(a_d^t\right)_{\mathrm{T}^s}^{\mathrm{T}^f} = \begin{cases} active & if \;\; p\left(a_d^t\right)_{\mathrm{T}^s}^{\mathrm{T}^f} \;\geq users \;\; threshold, F(t)_{\mathrm{T}^s}^{\mathrm{T}^f} \;\; is \;\; active \\ not\text{-}active & otherwise \end{cases} \tag{10}$$

A predator is said to be active and his/her motive of bullying during the interval $[\tau^s, \tau^f]$ for topic t if the probability of a predator participating in t, during that time period, exceeds the user-specified threshold, and $F(t)_{\mathrm{T}^s}^{\mathrm{T}^f}$ is active within that duration. The user enumerated threshold is calculated by taking an average of ϑ_a^t over predators for t. The contribution of a predator $a_{i,d}^t$ within $[\tau^s, \tau^f]$, using $P\left(a^{\mathrm{T}^s}|t\right) = \frac{p\left(a^{\mathrm{T}^s}|d^{\mathrm{T}^s}\right) \times p\left(t^{\mathrm{T}^s}|d^{\mathrm{T}^s}\right)}{p\left(d^{\mathrm{T}^s}\right)}$ per tome instance s, is mapped first in order to compute $p\left(a_{i,d}^t\right)_{\mathrm{T}^s}^{\mathrm{T}^f}$. Next, the total probability for predator a^t during $[\tau^s, \tau^f]$ is calculated as $\sum_{\mathrm{T}^s}^{\mathrm{T}^f} P\left(a^{\mathrm{T}^s}|t\right)$. **Figure 6** shows the activity of predators over time. For example, the activity of predators in bullying topic $t_{5,d5}$ during [15:00,21:00] can be analyzed in the following manner. Initially, the specified threshold is determined as 0.1770, for the average of ϑ_a^t. Then the mapping function is calculated for all predators. For example, a predator a_5 and time instance s = 15:00 are considered to analyze. The mapping function is calculated as $P(a_{5,\mathrm{T}}15:00|t_5) = 0.0547$ and then the total probability of a_5 is estimated by calculating $\sum_{\mathrm{T}^{15:00}}^{\mathrm{T}^{21:00}} P\left(a_{5,\mathrm{T}^s}|t_5\right) = 0.2307$. When applying the transition function $F\left(a_d^t\right)_{\mathrm{T}^s}^{\mathrm{T}^f}$, the predators ($a_1$,$a_3$) are active for bullying topic $t_{5,d1}$ and the predators (a_2,a_4,a_5) are not active.

3.4 Performance evaluation

The Perplexity of the model is used on test documents to estimate the execution of model and it is a customary measure for evaluating the operation of a probabilistic model. The adapted models are compared by means of perplexity on test datasets. Perplexity is extensively used in a probabilistic model for checking their quality. The perplexity of a couple of trial texts, (w_d,p_d) for d ϵ D^{test}, is characterized as the exponential of the negative standardized predictive likelihood underneath the representation,

$$perplexity(wd|pd) = \exp\left[-\frac{\ln p(wd|pd)}{Nd}\right] \tag{11}$$

MOTIVE = RACISM									
TOPIC 5		**TOPIC 10**		**TOPIC 15**		**TOPIC 20**		**TOPIC 25**	
EXTREMISM		**HOMOPHOBIA**		**VIOLENCE**		**REF. TO HANDICAPS**		**SLURS**	
Incorrect	0.0271	ColdSweat	0.0265	Shit	0.0752	Fuck	0.0798	Pussi	0.0321
Improper	0.0242	Dread	0.0254	Bullshit	0.0621	Ass	0.0767	Dog	0.0312
Indecent	0.0231	Fearful	0.0235	Piss	0.0506	Dumb	0.0722	Filthy	0.0304
Ineligible	0.0225	Horror	0.0223	Aggrieve	0.0254	Blind	0.0342	Crow	0.0294
Unfit	0.0214	Panic	0.0212	Tee toe	0.0232	Cracy	0.0212	Nitchie	0.0276
Unsuited	0.021	Phobia	0.0203	Nose	0.0215	Daft	0.0203	Peckerwood	0.0253
Room	0.0197	Scare	0.0194	Gotoofar	0.0201	Autism	0.0167	Cameljockey	0.0238
Raffish	0.0193	Terror	0.0187	Rufflesb's feathers	0.0176	Freak	0.0154	Nigger	0.0221
Square peg	0.0184	Alarm	0.0176	Aggravate	0.0154	Gimpy	0.0132	Peckerwood	0.0213
Unworthy	0.0173	Fright	0.0169	Burn	0.0132	Windowlicker	0.0121	Wigger	0.0201
Predators: Victims	Prob	Predators: Victims	Prob	Predators: Victims	Prob	Predators: Victims	Prob	Predators: Victims	Prob
P1: V1	0.0547	P1: V1	0.0341	P4: V4	0.0352	P3: V5	0.0421	P1: V6	0.0284
P2: V2	0.0367	P2: V2	0.0288	P1: V2	0.0254	P2: V6	0.0325	P5: V7	0.0257
P3: V3	0.0361	P3: V3	0.0254	P1: V3	0.0246	P1: V3	0.0208	P4: V5	0.0236
MOTIVE = SEXUAL									
TOPIC 30		**TOPIC 35**		**TOPIC 40**		**TOPIC 45**		**TOPIC 50**	
CRUDE LANGUAGE		**IMPLICIT LANGUAGE**		**INDECENT PROPOSALS**		**UNREFINED LANGUAGE**		**SLANG WORDS**	
Gay	0.0738	Dirty	0.0643	Mood	0.0547	Bitch	0.0705	Pull	0.0456
Suck	0.0711	Bed	0.0508	Lick	0.0519	Freak	0.0699	Bumpuglies	0.0423
Naked	0.0588	Frequent	0.0491	Kiss	0.0508	Fat	0.0663	Fug	0.0321

MOTIVE = SEXUAL									
TOPIC 30		**TOPIC 35**		**TOPIC 40**		**TOPIC 45**		**TOPIC 50**	
CRUDE LANGUAGE		**IMPLICIT LANGUAGE**		**INDECENT PROPOSALS**		**UNREFINED LANGUAGE**		**SLANG WORDS**	
Sexy	0.0569	Sleep	0.0282	Hangnow	0.0485	Happyhappy	0.0341	Randy	0.0307
Kickit	0.0445	Kneedeep	0.0241	Givebusiness	0.0465	Poundduck	0.0324	Juicy	0.0284
FuckforOL’	0.0432	Encounter	0.0215	Monkeylove	0.0328	Homerun	0.0307	Hempedup	0.0245
Getdown dirty	0.0421	Donasty	0.0208	Sexytime	0.0319	Smack	0.0284	Jiffystiffy	0.0209
Slap	0.0316	doublebag	0.0165	Intimacy	0.0206	Serve	0.0271	Ride	0.0154
Hump	0.0307	Giveitup	0.0154	Cottage	0.0191	Jellosex	0.0135	Smush	0.0124
Screw	0.0201	Getlucky	0.0142	Raunchy	0.0147	Score	0.0104	Trim	0.0107
Predators: Victims	Prob	Predators: Victims	Prob	Predators: Victims	Prob	Predators: Victims	Prob	Predators: Victims	Prob
P1: V1	0.0737	P4: V4	0.0541	P3: V5	0.0452	P1: V 6	0.0595	P3: V5	0.0354
P2: v2	0.0552	P1: V2	0.0428	P2: V6	0.0321	P5: V7	0.0467	P2: V6	0.0241
P3: V3	0.0324	P1: V3	0.0367	P1: V3	0.0276	P4: V5	0.0354	P1: V3	0.0211

MOTIVE = OUTRAGE		**MOTIVE = IRRELEVANT**		**MOTIVE = UNKNOWN**	
TOPIC 60		**TOPIC 70**		**TOPIC 90**	
ANGER		Make out	0.0267	Outhouse	0.0246
Bitterness	0.0365	Marquee	0.0235	Pant	0.0232
Hard	0.0354	Mate	0.0223	Pass out	0.0214
Storm	0.0321	Minor	0.0215	Patient	0.0208
Irritation	0.0306	Moot	0.0209	PC	0.0179
Wrath	0.0268	MP	0.0201	Period	0.0165
Fury	0.0251	MUM	0.0189	Plant	0.0152

MOTIVE = OUTRAGE		MOTIVE = IRRELEVANT		MOTIVE = UNKNOWN	
TOPIC 60		**TOPIC 70**		**TOPIC 90**	
ANGER					
Resent	0.0237	Nappy	0.0154	POP	0.0143
Rancor	0.0209	Natter	0.0142	Restroom	0.0137
Grudge	0.0192	Nick	0.0126	Rider	0.0129
Flap	0.0163	Nonce	0.0118	Sick	0.0109
Predators: Victims	Prob	Predators: Victims	Prob	Predators: Victims	Prob
P1: V7	0.0241	P5: V2	0.0207	P2: V6	0.0175
P2: V4	0.0219	P4: V5	0.0165	P5: V8	0.0154
P3: V3	0.0147	P1: V3	0.0126	P4: V5	0.0132

Table 3.
The distribution for the different bullying topics from the documents.

Better simplification functioning is designated by means of a lesser perplexity on a held-out document. The derivation of the likelihood of a collection of texts specified the predator is a uncomplicated computation in Bully-LDA model.

$$p(wd|pd) = \int d\theta \int d\phi p(\theta|Dtrain)p(\phi|Dtrain) * \prod_{m=1}^{Nd} \left[\frac{1}{Ad} \sum\nolimits_{i \in pd,j} \theta ij \phi wmj \right] \quad (12)$$

The term in the brackets is merely the probability for the word w_m specified the pair of predators p_d. The detailed results are exposed in **Figure 7**. These results

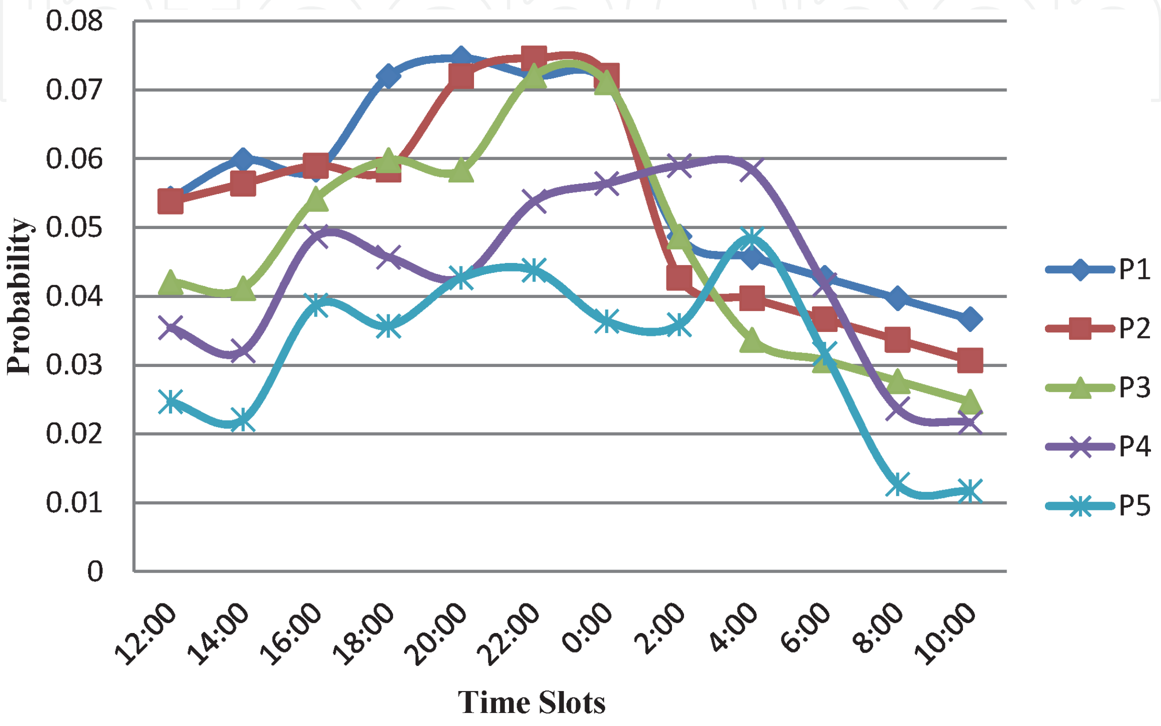


Figure 6.
Predators activity for bullying topic 30.

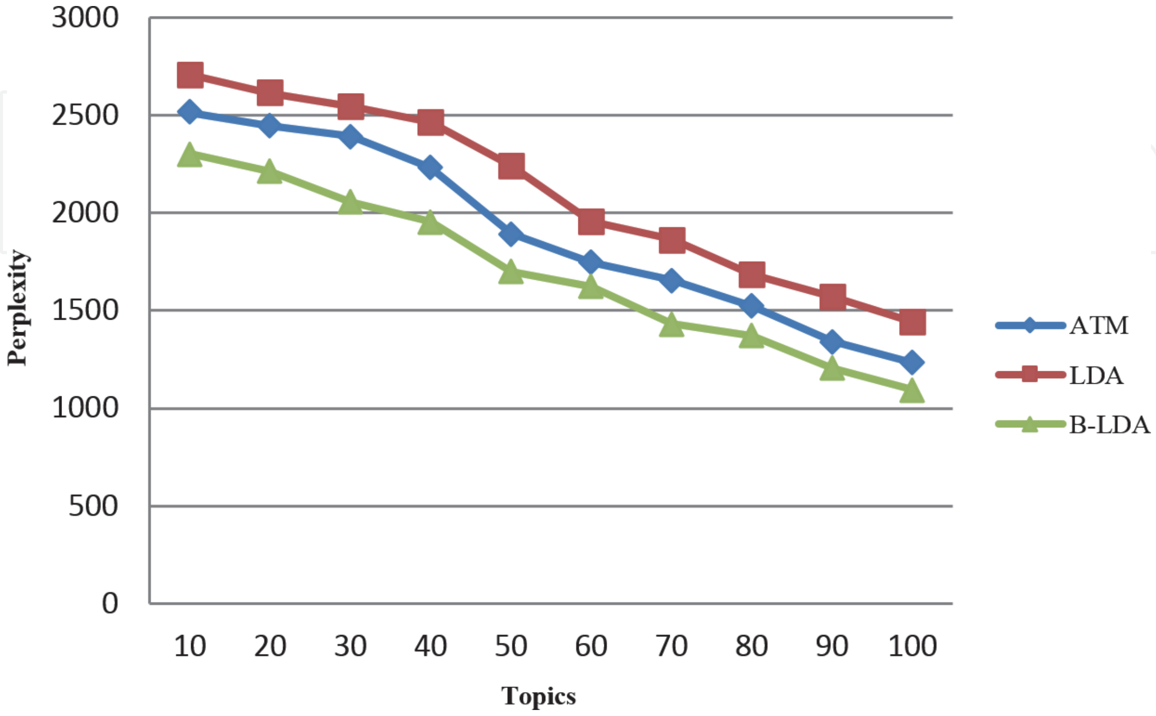


Figure 7.
Comparisons of different models in terms of perplexity.

indicate that B-LDA better generalizes performance than ATM and LDA. The improvement in generalization performance of B-LDA can be explained by its ability to better model when comparing with LDA and ATM model. If a word which has small probability in the bullying topics of training document, then it will cause an increase in perplexity. As the number of bullying topics increase, then the probabilities assigned to words get smaller in each bullying topic. Even though ATM models the roles of authors, does not show promising results and it is originally designed for the scenario where each document has multiple authors. It is clear that B-LDA achieves superior performances among all the adopted models. The perplexity of LDA, ATM, and B-LDA are closer and they decrease steadily with the increase of topics. According to human judgments, perplexity is not easy to correlate the results. So, it is necessary to compare the models using simple metrics like Precision, Recall, and F1 measure. The standard supervised classifier, i.e., Support Vector Machine (SVM), is adopted with B-LDA for classification. LibSVM was applied to the two-class classification problem using a linear kernel. Each post is an instance; positive classes contain bullying messages and negative classes contain non-bullying messages. A 10-fold cross-validation was performed in which the complete dataset was partitioned 10 times into 10 samples; in every round, nine portions were employed for exercising and the enduring section was applied for trial (**Figure 8**).

The functioning of the classifier was appraised on precision, recall and F-1 measure and these measures depend on the top-ranked features produced through B-LDA method against the truth set as tested on the datasets. Precision: The Aggregate number of accurately distinguished genuine harassing posts out of recovered tormenting cases. Recall: Number of effectively distinguished tormenting cases from an aggregate number of genuine harassing cases. F-1 measure: the equally weighted harmonic mean of precision and recall. **Table 4** shows the classifier performance.

3.5 Comparison of weighted B-TFIDF with baseline method

The weighted B-TFIDF method is compared with the work done in a content analysis in a web on four different datasets. The new feature selection method using weighted B-TFIDF proved that it is better than baseline. The outcomes are cataloged in **Table 5** and also indicate a very high precision, recall and F-1 measure on Twitter. In Kongregate precision fell down at the top 2000 features. In most of

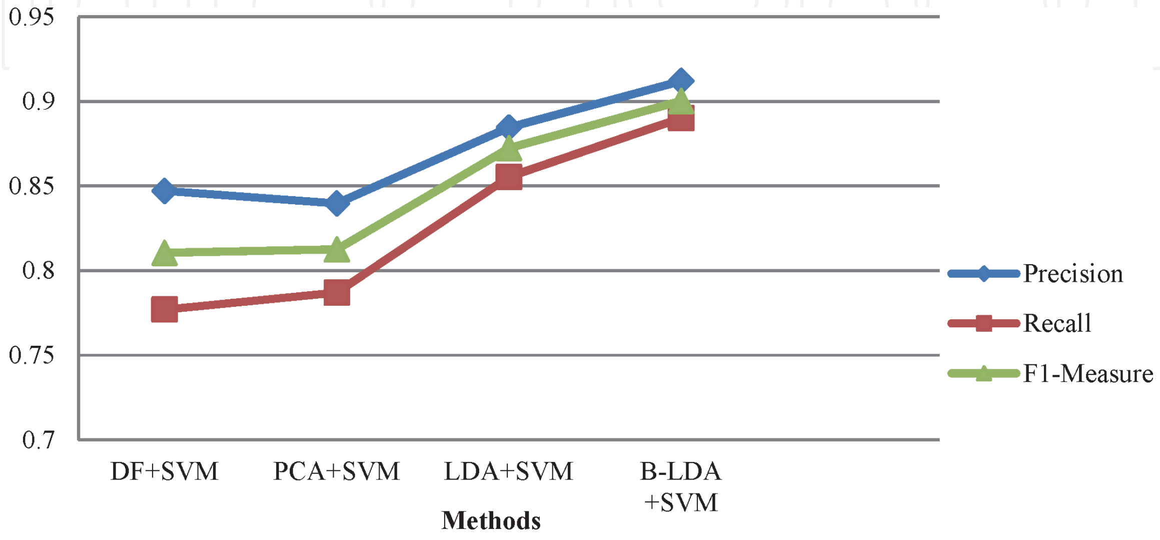


Figure 8.
Classifier performances based on different feature reduction methods.

the cases, the classifier performed almost similar, that is between 80 and 100%. On Myspace dataset recall is moderate nearing to 1. However, precision varies between 76 and 87% except at feature value 18,000 when it reaches 91%. Unlike other datasets, Slashdot performance is very low. Although recall is moderate, precision and F-1 measures decomposed while component set was low. Also, poor performance is observed at feature value 18,000. From this discussion, the performance of weighted B-TFIDF shows the best result (**Figure 9**).

3.6 Victim and predator identification

In order to identify cyber bullying predators and victims, there is need to determine the most active predators and the most attacked users. The most dynamic predators and victims, and look at the association of clients in a tormenting relationship as appeared in **Table 6** and it demonstrates that now and again there is more than one user at a similar rank. In this manner, users with a similar rank are gathered together. So it is important to notice that predators hailed at Rank I are additionally recognized as a victim at Rank II. Additionally, Rank II predators are Rank VII victims as well (**Figure 10**).

3.6.1 Graph representation

The major goal of a users' communication network are considered to identify predators and casualties. Gephi [15], a graphical interface is employed to monitor a user's link in the harassing posts in a network. **Figure 11** delineates the bullying network and it represents that a group of users obtained depend upon on the

Method	Precision	Recall	F-Measure
DF + SVM	0.8471	0.7770	0.8105
PCA + SVM	0.8397	0.7870	0.8125
LDA + SVM	0.8846	0.8554	0.8724
B-LDA + SVM	0.9121	0.8901	0.9003

Table 4.
Classifier performances based on different feature reduction methods.

		Kongregate	Slashdot	MySpace	Twitter
Baseline	Precision	0.35	0.32	0.42	0.62
Baseline	Recall	0.60	0.28	0.25	0.53
Baseline	F-1 measure	0.44	0.30	0.31	0.57
Weighted TFIDF	Precision	0.87	0.78	0.86	0.87
Weighted TFIDF	Recall	0.97	0.99	0.98	0.75
Weighted TFIDF	F-1 measure	0.92	0.87	0.92	0.81
Weighted B-TFIDF	Precision	0.95	0.96	0.96	0.98
Weighted B-TFIDF	Recall	0.93	0.84	0.93	0.96
Weighted B-TFIDF	F-1 measure	0.94	0.90	0.95	0.97

Table 5.
Comparison of weighted B-TFIDF with baseline method on other datasets.

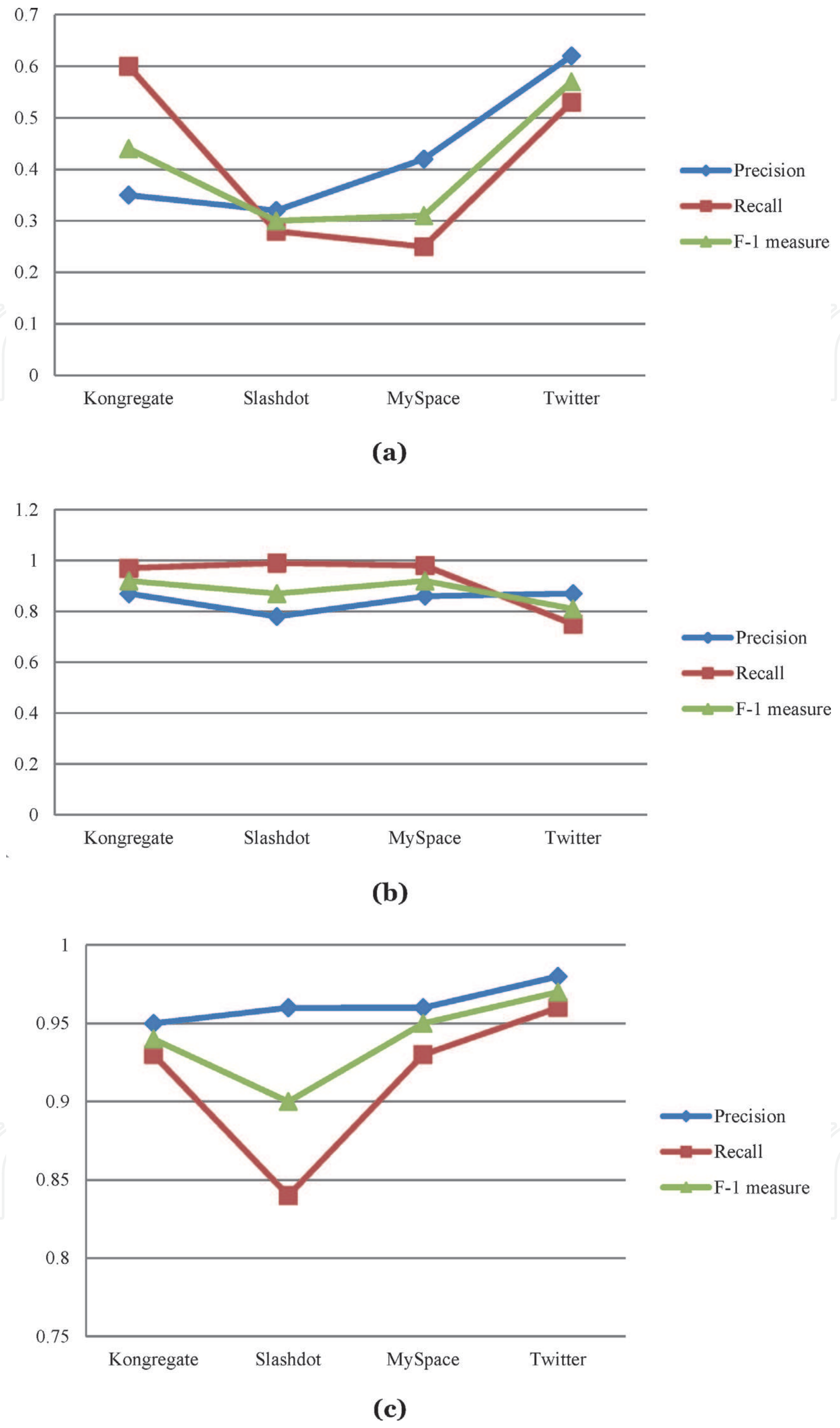


Figure 9.
(a) Base line method, (b) weighted TFIDF method, and (c) weighted B-TFIDF method.

tormenting messages by utilizing modularity theorem, in order to quantify the quality of segment of a system into sub-graphs or groups. Modularity is characterized as the summation of the weight of all the edges that sink inside the given subgroups less the expected part if edges were dispensed at arbitrary in a given graph.

Rank	I	II	III	IV	V	VI	VII	VIII
Number of users (predators)	4	2	1	1	2	7	3	2
Number of users (victims)	8	4	7	2	2	1	9	8

Table 6.
Performance of graph model: Predators and victims identification.

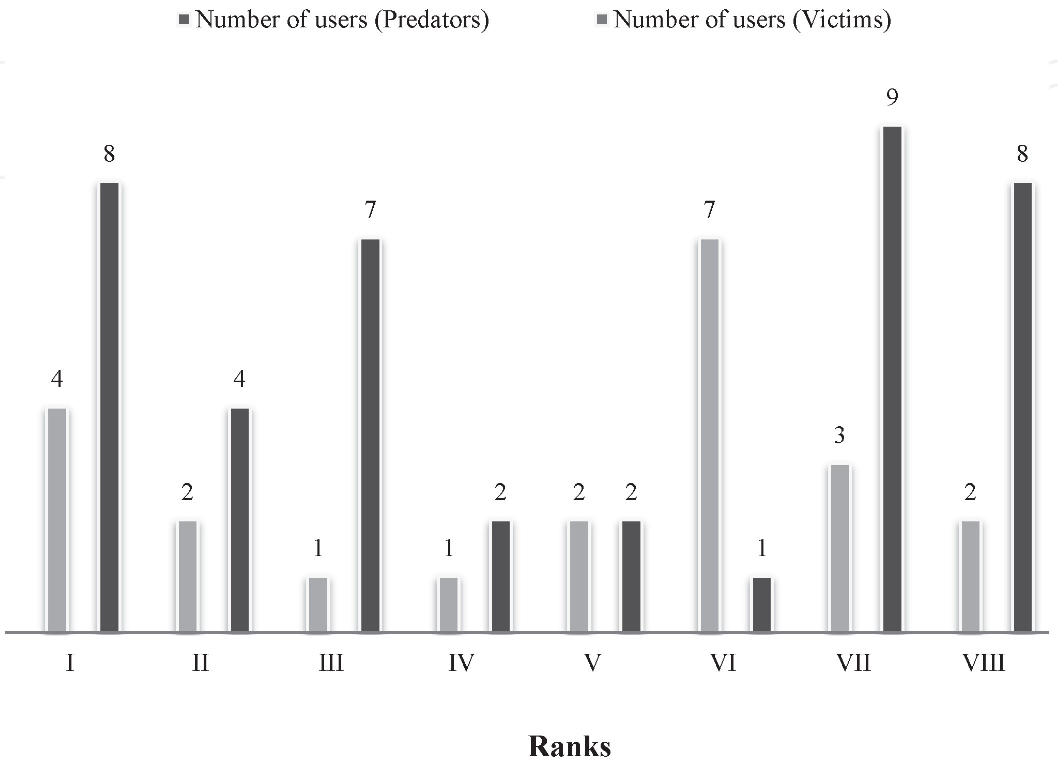


Figure 10.
Predators and victims identification.

As appeared in **Figure 11**, nine groups or communities, delineated by various colors are formed by considering users that are thickly connected inside the group contrasted with between group by utilizing modularity algorithm. The density of post indicates the badness embedded inside the post and it is calculated for each post. The thickness of a post is computed as the aggregate count of the harassing words within the post separated by the aggregate number of the words in the post. The HITS algorithm is utilized in order to recognize the predators and related casualties and it is also helpful to calculate their scores. The objective behind the HITS strategies is that in a network, the good hub pages point to good authorized pages which are connected by the good hub pages. The search query enters through web pages to recognize potential hub and authority pages with respect to the individual scores. Likewise, this concept is used to rank predators and casualties in a communication network.

Assumption: One bullying message is considered for each user.

Predator: Person who has posted at least one bullying message.

Victim: User who has received at least one bullying message.

Objective: To identify and to rank the most dynamic user as Predator and Victim.

Presently, a ranking method using the HITS module is utilized to detect predators and casualties. A user may be a predator and a victim depends upon on the

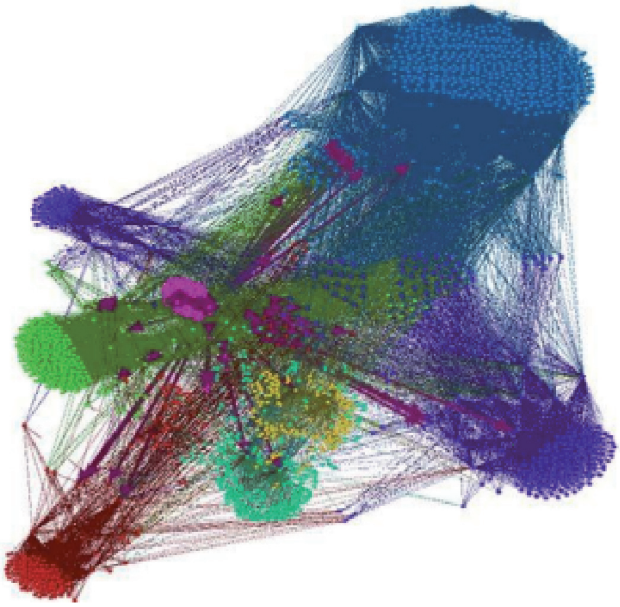


Figure 11.
Bullying network.

harassing messages he/she sends or receives. So, a user appointed as a predator and in addition with a casualty score. Predator and victim scores can be calculated by the following two equations.

$$p(u) \leftarrow \sum_{u \to y} v(y) \tag{13}$$

$$v(u) \leftarrow \sum_{y \to u} p(y) \tag{14}$$

Here, p(u) and v(u) are represented as the Predator and Victim scores respectively. u $\rightarrow y$ represents the existing harassing post from u to y, whereas $y \rightarrow u$ shows the presence of the bullying posting from y to u. The above equations are used for evaluating predator and casualty scores and also considered as repeatedly upgrade a set of equations. They depend upon the presumption that the most dynamic predator connects to the most dynamic victims by sending harassing posts. The most active victim is connected to the most dynamic predators by getting bullying messages. Basically, the user's predator score increases when the user (u) is connected with another user with a high victim score. In the same manner, the user's victim score increments when the user (u) is connected through received bullying messages to a user with a high predator score. The scores are computed through incoming degrees and outgoing degrees, and associated scores, in each and every iteration and this may give the result in large values. Subsequently, scores are standardized to unit length, i.e., each predator and victim scores is divided by the sum of all predator and victim scores respectively.

Then there is a necessity to define the ranking methods to the predators and victims which is depicted in the network diagram in **Figure 11**. In order to explain a real scenario in a simple manner, only five users are selected as depicted in **Figure 12** as an example and it depicts the recognition of the most dynamic predators and casualties in a bullying network. It is a weighted directed graph G = (U,A) with a set of nodes are represented as |U| and a set of arcs are represented as |A| where,

Each node u_i ϵ U is a user involved in the bullying conversation,

Each arc (u_i,u_j) ϵ A, is defined as a bullying message sent from u_i to u_j, The weight of arc (u_i,u_j), denoted as w_{ij}, is defined as a summation of in-degrees.

Predators and victims are recognized by the directed graph G with weight. The victim can be recognized with many incoming arcs and the predator can be recognized with many outgoing arcs of the respective nodes. This method is helpful to observe the most dynamic predator or a casualty.

3.6.2 Cyber bullying matrix

A cyber bullying matrix(w) is constructed to discover a predator and victim depends upon their individual scores. It is depicted in **Table 7**. It is formulated as a square adjacency matrix (it represents the incoming degrees and outgoing degrees of each node) of the subnet with entry w, which is a square adjacency grid of the sub collection with entry w_{ij}, where,

$$w_{ij} = \{\text{n if there be n harassing posts from } u_i \text{ to } u_j, 0 \text{ otherwise}\} \tag{15}$$

Since each client will have a casualty as well as a predator score, scores are represented as the vectors of n*1 dimension where i^{th} coordinate of the vector represent both the scores of the i^{th} user, say p_i and v_i respectively. To calculate scores, equations p(u) and v(u) are shortened as the casualty and predator renovating matrix–vector multiplication equations. For the preliminary iteration, p_i and v_i are started at 1. For every client (say, i = 1 to N) predator and victim notches are as follows:

$$p(ui) = wi1v1 + wi2v2 + ... + wiNvN \tag{16}$$

$$v(ui) = wi1p1 + wi2p2 + ... + wiNpN \tag{17}$$

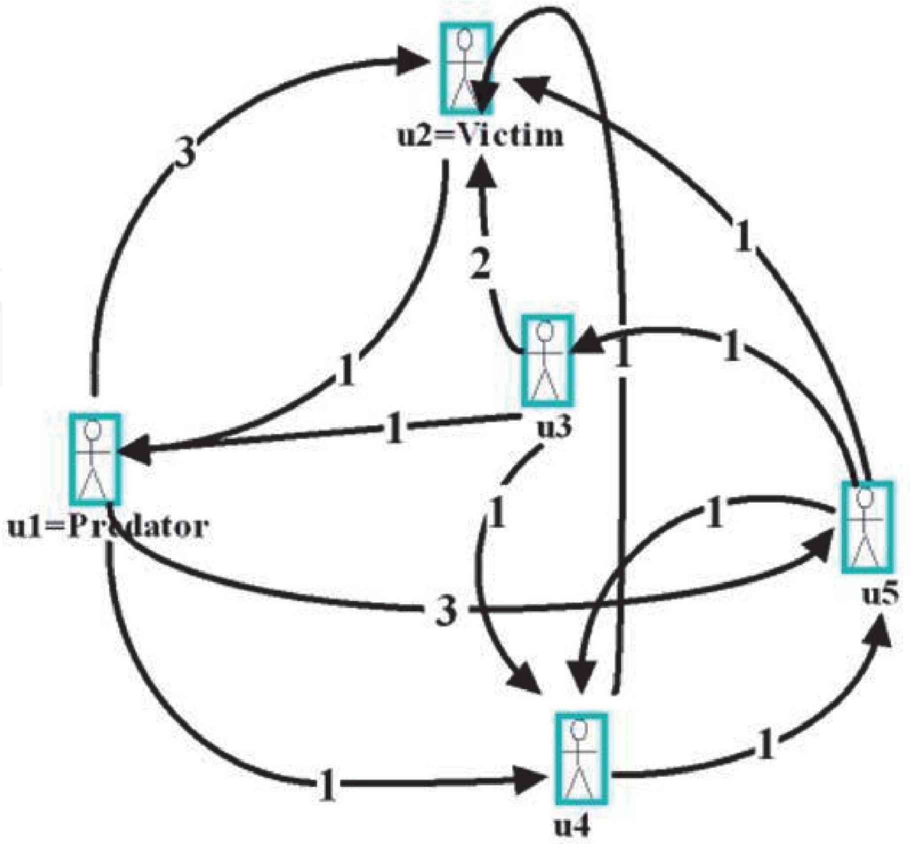


Figure 12.
Communication paths between predator and casualty.

Sender	**Recipient**						
	U_1	**U_2**	**U_3**	**U_4**	**U_5**	**… …**	**U_N**
U_1	0	3	0	1	3	… .	… .
U_2	1	0	0	0	0	… .	… .
U_3	1	2	0	1	0	… .	… .
U_4	0	1	0	0	1	… .	… .
U_5	0	1	1	1	0	… .	… .
… ..	… .	… .	… .	… .	… .	… .	… .
U_N	… .	… .	… .	… .	… .	… .	… .

Table 7.
Cyber bullying matrix (W).

When these equations congregate at a stable value (say k), it offers the final predator and casualty vector of each user. At last, to compute the eigenvector to acquire the predator and casualty scores.

Algorithm 1 gives a general framework of identification of the top-ranked most active predators and victims. In the algorithm N is a total number of users and Top is a threshold value, which is set manually.

Algorithm 1. Predators and casualty recognition.

```
Input: Set of consumer engaged in the chat with harassing
post, N, Top.
Output: Set of Top Casualty and Top Predator

    1. Take out dispatchers and receivers from N;
    2. Initialize predator and casualty vector each N;
    3. Generate adjacent matrix w using formula (15);
    4. Compute Predator and casualty vectors with iterative updating
       Eqs. (16) and (17), and normalize, until congregate at secure value k;
    5. Compute Eigen vectors to locate Predator and Casualty scores;
    6. Revisit high ranked Predators and Casualties.
```

4. Summary

The new system is achieved by two commitments. First, a Novel Statistical Application, which is established on the new Bully-LDA with the weighted B-TFIDF strategy on bullying like attributes. It also efficiently and effectively finds latent bullying features to cultivate the accomplishment of the classifier and also to reduce the feature sparsity. Secondly, a Graph Model lends a hand to pinpoint the attackers and causalities in social networks. Such a system would encompass the following function: Tweets Crawling, Tweet Preprocessing and Tokenization, Feature extraction and Frequency extraction, Text Representation Model, Text Classification, Category of Texts, Performance Evaluation, and Results.

The Twitter corpus consists of text communications by way of metadata such user ID, dispatching time, etc. Tweets Crawling is performed using many classes and techniques in order to get the information of the users' connected data and the details of the Tweets' which is done using Twitter's Application programming interface called "Twitter4j-core-4.02.jar." Tweets are shown in entirely colloquial manner, with more amount noise and variation in linguistics. For example, tweets contain a hefty quantity of novel words, interjections, repetitions, short words such as acronyms, words with missing letters, words with phonetic spelling like *Gud* for *Good*, etc. and also missing blank spaces between the words, such as *whatareyoudoing*, which increases the tweet length. All these things impose a huge burden in the analysis of the text. Text preprocessing module contains word segmentation, word processing, and subsequent analytical steps include like converting uppercase letters to lower case, stemming, eradicating stop words, superfluous characters and hyperlinks.

The proposed framework utilizing Bully-Latent Dirichlet Allocation through Support Vector Machine has been examined with Twitter messages. This system is based on a novel concept of applying text mining techniques to tweets for detecting Bullying messages and also to identify Predators. The weighted B-TFIDF function is used to enhance the execution of classification, in which bullying-like features are measured. The overall results using Bully-LDA + SVM and weighted B-TFIDF outperformed other models. This model has numerous benefits adding more accuracy, superior noise diminution, faster speed and greater automation. The results obtained were analyzed properly using different metrics. A range of performance measures for instance accuracy, recall and F1 measures were calculated. The analysis of results plainly displays that the system yields effective results in identifying bullying messages in a successful manner.

In this research, a methodology for cyber bullying recognition of the most operative predators and casualties are done powerfully and fruitfully. This chapter presents a framework for detecting cyber bullying in Twitter using Bully-Latent Dirichlet Allocation with support vector machine. The preprocessing procedures have pertained to tweets. First Bully-LDA, a statistical topic modeling is used on a massive Twitter Corpus, with the help of weighted B-TFIDF scheme to detect offensive words in tweets. Next, a graph representation is utilized to recognize the predators and casualties in Twitter.

Author details

K. Nalini[1]* and L. Jabasheela[2]

1 Bharathiyar University, India

2 Panimalar Engineering College, India

*Address all correspondence to: immanuelsamen@rediffmail.com

References

[1] Ibn Rafiq R, Hosseinmardi H, Han R, Mishra S, Lv Q. Scalable and Timely Detection of Cyber Bullying in Online Social Networks. France: ACM; 2018. pp. 1738-1747

[2] Zhong A, Li H, Squicciarini A, Rajtmajer S, Griffin C, Miller D, et al. Content-driven detection of cyberbullying on the Instagram social network. In: Proceedings of the Twenty-Fifth International Journal Conference on Artificial Intelligence, (IJCAI-16). 2016

[3] Sherly TT, Rosiline JB. Supervised feature selection based extreme learning machine (SFS-ELM) classifier for cyberbullying detection in Twitter. International Journal of Scientific and Research Publications. 2017;**7**(7): 367-373

[4] Gotardo MA. Topic modelling of online child pornography documents. International Journal of Social Science and Economic Research. 2018;**3**(2): 505-521

[5] Priyangika S, Jayalal S. Detection of cyberbullying on social media networks. In: International Research Symposium on Pure and Applied Sciences. 2016

[6] Kansara KB, Shekokar NM. A framework for cyberbullying detection in social network. International Journal of Current Engineering and Technology. 2015;**5**(1):494-498

[7] Parapar J, Losada DE, Barreiro Á. Combining psycho-linguistic, content based and chat-based features to detect predation in chat rooms. Journal of Universal Computer Science. 2014; **20**(2):213-239

[8] Huang Q, Singh VK, Atrey PK. Cyberbullying detection using social and textual analysis. In: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, ACM. 2014. pp. 3-6

[9] Zhao R, Zhou A, Mao K. Automatic detection of cyberbullying on social networks based on bullying features. In: Proceedings of the 17th International Conference on Distributed Computing and Networking, ACM. 2016. p. 43

[10] Chen Y, Zhou Y, Zhu S, Xu H. Detecting offensive language in social media to protect adolescent online safety. In: Privacy, Security, Risk and Trust (PASSAT), International Conference on Social Computing (Social Com), IEEE. 2012. pp. 71-80

[11] Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author topic models for information discovery. In: Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining, ACM Press, New York. 2004. pp. 306-315

[12] Blei D, Gri T, Jordan M, Tenenbaum J. Hierarchical topic models and the nested Chinese restaurant process. In: Seventeenth Annual Conference on Neural Information Processing Systems, NIPS. 2003

[13] Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America. 2004;**101**(1): 5228-5235

[14] Minka T, Lafferty J. Expectation-propagation for the generative aspect model. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. 2002. pp. 352-359

[15] Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks. In: International AAAI Conference on Weblogs and Social Media. 2009. pp. 361-362