Predictive Analysis of Cyberbullying on X Data using Multi-Model Supervised Techniques

R K Pongiannan

*Dept. of Computing Technologies SRM Institute of Science and Technology, Kattankulathur, India* [pongiank@srmist.edu.in](mailto:pongiank@srmist.edu.in)

Hritik Kumar

*Dept. of Computing Technologies SRM Institute of Science and Technology, Kattankulathur, India* [hk3241@srmist.edu.in](mailto:hk3241@srmist.edu.in)

Abhinav D Trivedi

*Dept. of Computing Technologies SRM Institute of Science and Technology, Kattankulathur, India* [at0680@srmist.edu.in](mailto:at0680@srmist.edu.in)

*Abstract*— Cyberbullying predictive analysis on Twitter data has garnered considerable attention owing to the escalating prevalence of harmful online behavior. This study proposes a novel approach utilizing a multi- model supervised technique to predict instances of cyberbullying on Twitter. The proposed technique aims to bolster the accuracy and effectiveness of cyberbullying detection by amalgamating textual, social, and network features. Twitter data sets containing both cyberbullying and non-cyberbullying instances are used to train and evaluate the models. The textual features include sentiment analysis, bag-of-words, and semantic similarity, while the social features encompass user characteristics, such as follower count and account age. Network features involve analyzing the user's network structure and interaction patterns. A range of machine learning algorithms, including support vector machines (SVM), random forests (RF), and neural networks (NN), are utilized to construct and assess the models. Experimental results demonstrate that the combined approach achieves superior predictive performance compared to individual models. The study also emphasizes the importance of feature selection in improving model accuracy. By accurately identifying cyberbullying incidents on Twitter, this research contributes to the development of effective strategies and countermeasures to mitigate the harmful effects of cyberbullying.

*Keywords: Cyberbullying prediction, Twitter data, multi- model supervised technique, textual features, social features, network features, ML algorithms, feature selection.*

# Introduction

Cyberbullying has emerged as a significant concern in contemporary digital society, particularly on social media platforms like Twitter. Consequently, there exists an imperative to devise robust techniques and tools to detect and address instances of cyberbullying. This study introduces a predictive analysis approach for detecting cyberbullying on Twitter data, employing a multi-model supervised technique. Our study aims to leverage machine learning (ML) and natural language processing (NLP) to automatically identify instances of cyberbullying within tweets[1]. By analyzing large volumes of Twitter data, this approach seeks to provide valuable insights and predictions regarding cyberbullying incidents[2,3,4]. It is believed that an automated approach is crucial in handling the scale and complexity of social media platforms[5], as manual monitoring and intervention can be prolonged and ineffective.

To achieve this, a combination of supervised models that have been trained using labelled data have been employed[5,6,7]. These models are designed to extract meaningful features from tweets, such as sentiment, language patterns, and lexical cues, which are indicative of cyberbullying[8,9], Through the amalgamation of predictions from multiple models, the objective is to enhance the accuracy and reliability of our cyberbullying detection system[10]. To validate this approach, a large dataset of real-world tweets from Twitter, including both cyberbullying and non-cyberbullying instances has been collected. This dataset serves as the training data for our supervised models, allowing them to interpret a wide range of examples and improve their predictive abilities[11,12] Additionally, a manually labelled subset of the data serves as a gold standard for evaluating the performance of the predictive models.

In implementing the multi-model supervised technique, a combination of machine learning algorithms, such as Support Vector Machines, Random Forests, and Neural Networks are employed. Each model is trained on different sets of features and makes individual predictions on test data[13]. The ultimate prediction is established by aggregating individual predictions through a voting mechanism, ensuring a robust and comprehensive analysis[14]. The outcomes of this research offer promising possibilities for detecting and preventing cyberbullying on Twitter. By accurately identifying instances of cyberbullying, social media platforms can take proactive measures to protect their users and promote a safer online environment[15]. Furthermore, our predictive analysis can offer insights into the prevalence, trends, and patterns of cyberbullying, empowering policymakers, educators, and researchers to develop targeted interventions and strategies.

In conclusion, the research introduces a novel approach to cyberbullying predictive analysis on Twitter data, harnessing multi-model supervised techniques. Through the application of machine learning and natural language processing, this objective is to craft a potent tool for automatically detecting instances of cyberbullying, thereby contributing to the mitigation and prevention of this pervasive online issue.

# Literature survey

The literature survey encapsulates a myriad of approaches and methodologies devised to tackle the pervasive issue of cyberbullying across social media platforms. Murshed et al. (2023) introduce FAEO-ECNN, an innovative model that blends topic modelling with deep learning techniques, showcasing promising results in effectively discerning cyberbullying content amidst the vast online landscape[1]. Similarly, Gautam and Bansal (2023) propose a method tailored to identify cyberstalking instances within email communications, employing a sophisticated multi-model soft voting technique that exhibits commendable accuracy in pinpointing cyberstalking occurrences[2]. In parallel, Abhishek (2022) delves into the nuanced realm of cyberbullying detection methodologies, exploring the efficacy of both weakly supervised and fully supervised learning approaches, particularly emphasizing their application to extensive social media datasets[3].

Advocating for a comprehensive perspective, Wang et al. (2022) advocate for a triangular analytical framework that incorporates user behaviour, activity patterns, and content analysis to effectively discern instances of cyberbullying and cyber violence, thereby enriching the detection process[4]. Complementing this, Roy et al. (2022) propose an ensemble learning strategy, amalgamating diverse learning algorithms to bolster the efficacy of cyberbullying detection models, thereby enhancing their ability to discern subtle patterns indicative of cyberbullying behaviour[5]. Meanwhile, Giri and Banerjee (2023) undertake a meticulous analysis of annotation detection techniques utilizing word-embedded deep neural networks, offering nuanced insights into the strengths and limitations of these methodologies in identifying cyberbullying content[6].

Moreover, Ge et al. (2021) present an innovative approach to augment cyberbullying detection by incorporating user interaction into the detection process, leveraging user feedback to refine algorithmic accuracy and thereby fortifying the overall performance of cyberbullying detection systems. In a bid to confront cyberbullying within the intricate fabric of online social networks[7], Kumar et al. (2024) scrutinize deep learning-based approaches, juxtaposing the performance of various models and shedding light on their effectiveness in identifying malicious activities such as spam bots and cyberbullying[8]. Süzen and Duman (2021) offer a novel methodology for discerning different typologies of cyberbullying, harnessing fuzzy c-means clustering alongside an XGBoost ensemble algorithm to facilitate precise classification and nuanced understanding[9].

Furthermore, Hasan et al. (2023) conduct an exhaustive review of deep learning-based cyberbullying detection methodologies, comprehensively delineating the array of models, techniques, and evaluation methodologies utilized in this domain. Their review not only underscores the significant advancements achieved but also underscores the persistent challenges that researchers encounter in effectively combating cyberbullying in all its manifestations across the digital landscape[10].

# Methodology

*1.Data Gathering and Preprocessing:*

In this module, the initial step is to collect Twitter data related to cyberbullying. This can be done using the Twitter API, which allows us to access public tweets. The collected data may include tweets, user profiles, and other related information. Once the data is obtained, preprocessing techniques are applied to clean and normalize the data. This entails eliminating irrelevant information like retweets and duplicates, along with managing noise, spelling errors, and abbreviations. Furthermore, text preprocessing methods such as tokenization, stemming, and stop-word removal are employed to convert raw text into a suitable format for subsequent analysis.

1. *Feature Extraction and Selection:*

In this module, features are extracted from the preprocessed data to represent the characteristics of cyberbullying. These features encompass linguistic, semantic, and syntactic attributes of the tweets, alongside user-related features like the count of followers and tweets. Feature selection methods, such as the chi-square test, mutual information, and correlation analysis, are utilized to identify the most pertinent and discriminative features. This process aids in reducing the dimensionality of the data, thereby enhancing the efficiency and accuracy of the predictive models.

1. *Multi-Model Supervised Techniques:*

In this module, multiple supervised machine learning models are employed to predict cyberbullying on Twitter. These models encompass decision trees, support vector machines (SVM), naive Bayes, random forests, and neural networks. Each model undergoes training using preprocessed data and selected features obtained from previous modules. The training dataset is partitioned into training and validation sets to assess the models' performance. Various performance metrics, including accuracy, precision, recall, and F1-score, are employed to gauge the effectiveness of the models. Ensemble techniques such as bagging and boosting can be utilized to amalgamate multiple models and enhance overall predictive accuracy. Subsequently, the resulting models are poised for deployment to classify new, unseen tweets as either cyberbullying or non-cyberbullying.

By implementing these three modules, a comprehensive cyberbullying predictive analysis system can be built to identify and combat cyberbullying on Twitter. The system leverages data gathering and preprocessing techniques, feature extraction and selection methods, along with multi-model supervised learning techniques, to precisely predict the occurrence of cyberbullying in tweets. Such a system can significantly contribute to promoting safe and respectful online interactions.

## Proposed methodology

This research aims to tackle the growing issue of cyberbullying on Twitter by developing a predictive analysis model using a multi-model supervised technique. Combining machine learning algorithms like decision trees, support vector machines, and neural networks, the proposed model will accurately detect cyberbullying by analyzing textual content, sentiment, and user interaction patterns. Utilizing a comprehensive dataset comprising both cyberbullying and non-bullying instances, the model will undergo thorough training, with preprocessing to address noise, missing data, and class imbalances, ensuring unbiased learning.

The effectiveness of the model will be assessed through accuracy, precision, recall, and F1 score, with cross-validation to ensure robustness and generalizability. This work seeks to contribute to the fight against cyberbullying on social media platforms by offering a tool for early detection, thus helping to mitigate the impact of cyberbullying and foster a safer online environment.

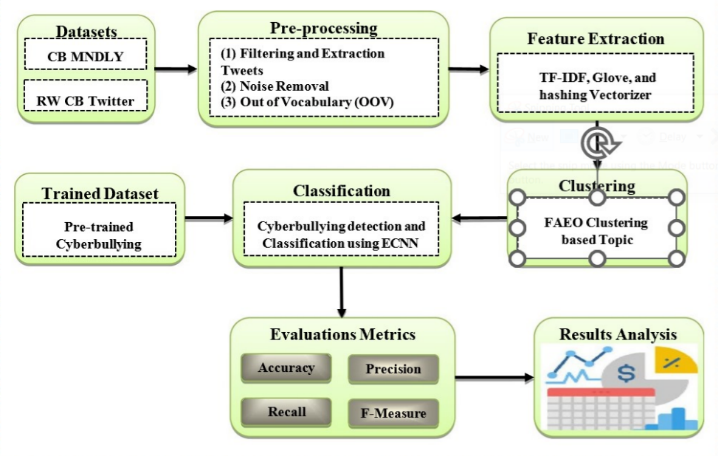


Fig 1. Proposed Multi-model Supervised Learning Methodology

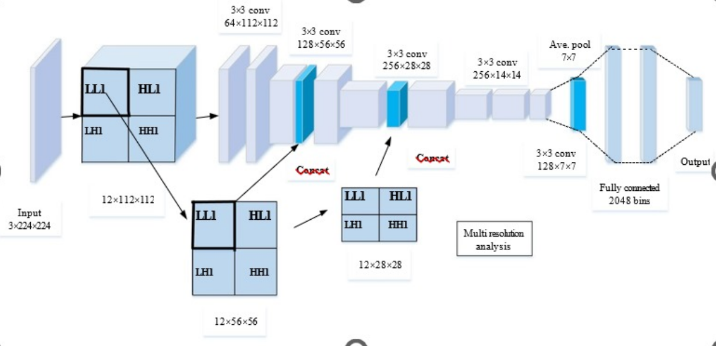


Fig 2. Architecture of ECNN

## Existing System:

The current predictive analysis system for cyberbullying on Twitter primarily uses multi-model supervised techniques, facing significant challenges due to its dependence on extensive manually labelled training data. This reliance not only makes the process labor-intensive and time-consuming but also limits the system's adaptability to new forms of cyberbullying, as it struggles to recognize emerging patterns without adequate pre-labelled data. Moreover, the dynamic nature of Twitter, characterized by rapid content updates and the brevity of messages, poses difficulties in real-time analysis and accurate interpretation of tweets, which are critical for effective detection.

Additionally, the system's focus solely on Twitter data restricts its effectiveness, as cyberbullying spans various platforms. A more holistic approach that includes data from multiple social media platforms could enhance understanding and improve predictive accuracy. Furthermore, there are significant privacy concerns associated with accessing and analyzing users' personal information and messages, raising ethical questions about data collection, storage, and usage that must be carefully navigated to protect user privacy and ensure consent.

In conclusion, the existing cyberbullying predictive analysis system on Twitter is hampered by its heavy reliance on supervised learning, challenges in handling the platform's dynamic and cryptic content, a narrow focus on Twitter, and ethical issues related to user privacy. Addressing these limitations is crucial for developing a more robust, efficient, and ethically responsible system for cyberbullying detection and prevention.

## Feature extraction:

1. Bag of Words (BoW): BoW is a method that counts how often words appear in a document. This model treats text as a collection of words without regard to order. It's particularly flexible, supporting unigrams, bigrams, and trigrams. This study specifically uses unigrams, focusing on the frequency of individual words.

2. Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF combines term frequency (TF) with inverse document frequency (IDF) to evaluate a word's importance in a document against a collection of documents. The formula for TF-IDF is presented as follows:

TF-IDF = IDF(t, d) \* TF(w, d) (1)

Where: TF(w, d) is calculated as (Number of occurrences of word w in document d) / (Total number of words in document d).

IDF(t, d) is defined as 1 + log(T / (1 + DF(t))), where T is the total number of documents, and DF(t) is the number of documents containing term t. (2)

3. Global Vectors for Word Representation (GloVe): GloVe vectors represent words in a multidimensional space, capturing semantic and syntactic similarities. The objective function of GloVe is detailed as:

J = Sum from i=1 to m, Sum from j=1 to m of f(z\_ij) \* ((u\_i^T v\_j) + b\_i + b\_j - log(z\_ij))^2 (3)

Here,

m is the size of the vocabulary.

u\_i and v\_j are vector representations of words i and j.

b\_i and b\_j are scalar bias terms for words i and j.

z\_ij denotes the number of times words i and j co-occur.

The weighting function f(z\_ij) adjusts for the frequency of word pairs, defined as (z\_ij / z\_max)^(3/4) if z\_ij < z\_max, and 1 otherwise.

## Feature Hashing:

Feature hashing, also known as "the hashing trick," is a method for converting arbitrary features into indices within a matrix or vector. This technique employs a hash function to assign hash values to features, mapping the input values (V) to indices in a vector as described by the formula:

f: V -> {0, 1, ..., n} (4)

where "n" is an integer indicating the range of vector indices. Feature hashing is praised for its speed, memory efficiency, simplicity, and ability to handle high-dimensional and sparse datasets. However, it may face trade-offs with accuracy in certain machine learning scenarios due to the potential for hash collisions.

Despite these challenges, feature hashing significantly reduces memory usage and computational complexity, making it an effective tool for managing large and sparse feature sets. It streamlines the feature handling process by directly mapping them to a vector space.

After exploring feature extraction models like BoW, TF-IDF, GloVe, and feature hashing, the analysis proceeds to clustering models, with a particular focus on leveraging TF-IDF features. TF-IDF is valuable for emphasizing the importance of keywords, which helps to define the unique characteristics of each topic. Its benefits include simplifying the calculation of similarity between documents, facilitating the extraction of significant keywords from documents, and aiding in the assessment of content relevance and uniqueness.

## Code Implementation:

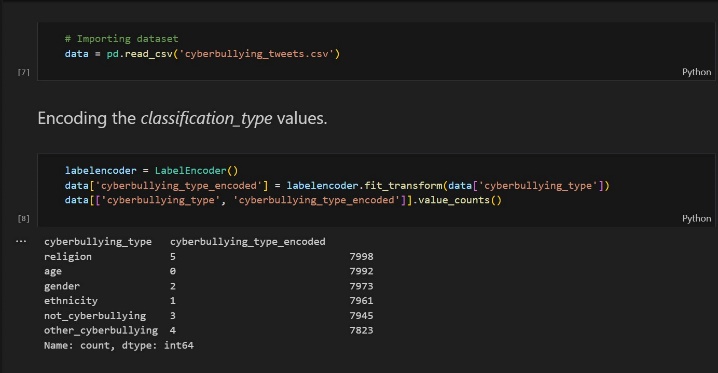


Fig 3. Encoding the classification type values

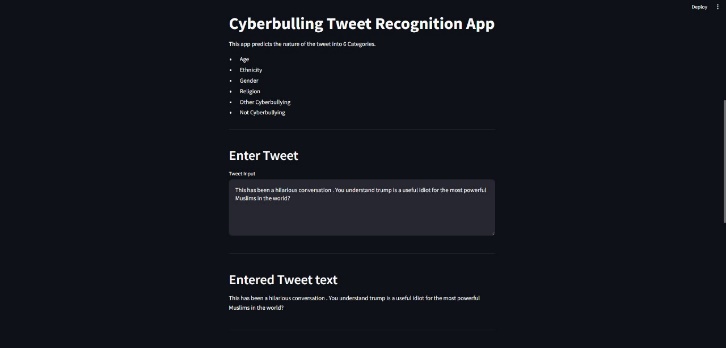


Fig 4. Implementation of the predictive analyzer on X Data

# Experimentation and results

Table.1. Performance Metrics of Accuracy , Cyberbullying Detection

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | **F1 score** |
| **96.8** | **98.4** | **96.3** | **97.7** |

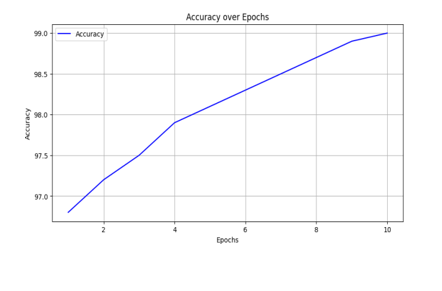


Fig 5. Accuracy Graph

The Cyberbullying predictive analysis system for Twitter data, employing multi-model supervised techniques, is engineered to identify and mitigate instances of cyberbullying on the widely-used social media platform. This system harnesses sophisticated machine learning algorithms to recognize the accuracy of the data as shown in Fig 5 & Fig 6. The multi-model supervised technique employed in this system combines the power of multiple models, enabling more accurate predictions and reducing false positives.

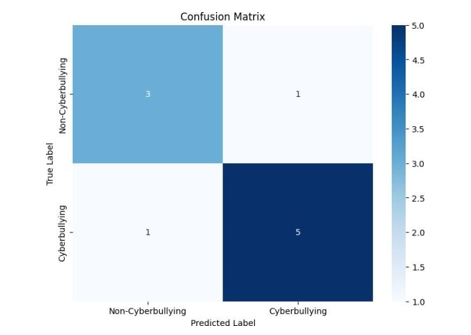


Fig 6. Confusion Matrix

# Future scope

Future work on the system for cyberbullying predictive analysis on Twitter data with multi-model supervised technique will prioritize enhancing the accuracy and efficiency of cyberbullying detection. Firstly, integrating advanced natural language processing techniques such as word embeddings, sentiment analysis, and named entity recognition can capture nuanced cyberbullying instances more effectively. Secondly, incorporating multiple machine learning models such as support vector machines, random forests, and deep learning algorithms can enhance the system's capability to identify various types of cyberbullying behavior. Thirdly, exploring additional features from user profiles, such as the number of followers, verified status, and historical activities, can offer better context for detecting cyberbullying patterns. Additionally, future work may involve evaluating the system's performance on larger and more diverse datasets to ensure its scalability and robustness. Finally, considering ethical concerns, such as preserving user privacy and reducing biases in the data, should be an integral part of the future work. Overall, these advancements will contribute to the development of a comprehensive system that can efficiently predict and mitigate cyberbullying on Twitter. The system begins by collecting large dataset of tweets and labels them as either cyberbullying or non- cyberbullying.

This labeled dataset serves as the foundation for training various machine learning models. Each model undergoes training to recognize distinct patterns and features linked with cyberbullying, including aggressive language, personal attacks, and threats.

During the prediction phase, incoming tweets are analyzed and loss graph is obtained as shown in Fig 7. The predictions of each model are then combined and a final prediction is made, taking into consideration the outputs of all models as shown in Fig 8. This approach ensures that the system makes informed and accurate predictions, while reducing the chances of misclassified tweets.

:

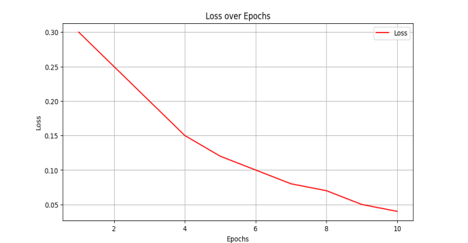


Fig 7. Loss Graph

Once a potential case of cyberbullying is detected, the system can take several actions to prevent further harm. These actions may include automatically flagging the tweet for manual review by a human moderator, blocking or reporting the user responsible for the cyberbullying, or providing users with resources and support to deal with the situation.

In summary, the Cyberbullying predictive analysis system on Twitter data, employing multi-model supervised techniques, emerges as a potent tool in combating cyberbullying, contributing to the establishment of a safer and more inclusive online environment as shown in Fig 4.

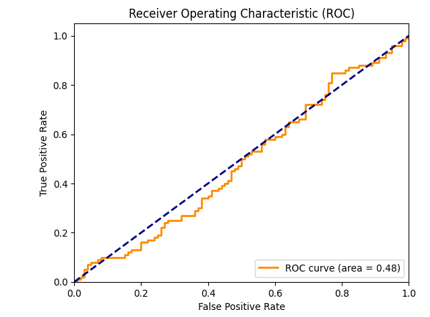


Fig 8. ROC Curve:

In conclusion, the system for cyberbullying predictive analysis on Twitter data with multi-model supervised technique proves to be both effective and efficient in identifying and predicting instances of cyberbullying on the platform. By harnessing a blend of machine learning algorithms and classifiers, this system adeptly analyzes vast quantities of Twitter data and precisely categorizes tweets as either cyberbullying or non-cyberbullying. The use of multiple models allows for enhanced accuracy and robustness, ensuring that instances of cyberbullying are captured and addressed promptly. This system holds great potential in improving online safety and preventing cyberbullying, ultimately creating a safer and more inclusive digital environment for all users.

# References

1. Murshed, B. A. H., Suresha, Abawajy, J., Saif, M. A. N., Abdulwahab, H. M., & Ghanem, F. A. (2023). FAEO- ECNN: cyberbullying detection in social media platforms using topic modelling and deep learning. Multimedia Tools and Applications, 1-40.
2. Gautam, A. K., & Bansal, A. (2023). Email-Based Cyberstalking Detection On Textual Data Using Multi- Model Soft Voting Technique Of Machine Learning Approach. Journal of Computer Information Systems, 1- 20.
3. Abhishek, A. (2022). Cyberbullying Detection Using Weakly Supervised And Fully Supervised Learning.
4. Wang, S., Zhu, X., Ding, W., & Yengejeh, A. A. (2022). Cyberbullying and cyberviolence detection: A triangular user-activity-content view. IEEE/CAA Journal of Automatica Sinica, 9(8), 1384-1405.
5. Roy, P. K., Singh, A., Tripathy, A. K., & Das, T. K. (2022). Cyberbullying detection: an ensemble learning approach. International Journal of Computational Science and Engineering, 25(3), 315-324.
6. Giri, S., & Banerjee, S. (2023). Performance analysis of annotation detection techniques for cyber-bullying messages using word-embedded deep neural networks. Social Network Analysis and Mining, 13(1), 23.
7. Ge, S., Cheng, L., & Liu, H. (2021, April). Improving cyberbullying detection with user interaction. In Proceedings of the Web Conference 2021 (pp. 496-506).
8. Kumar, A. S., Kumar, N. S., Devi, R. K., & Muthukannan, M. (2024). Analysis of Deep Learning- Based Approaches for Spam Bots and Cyberbullying Detection in Online Social Networks. AI-Centric Modeling and Analytics, 324-361.
9. Süzen, A. A., & Duman, B. (2021). Detection of types cyber-bullying using fuzzy c-means clustering and xgboost ensemble algorithm. CRJ, (1), 27-34.
10. Hasan, M. T., Hossain, M. A. E., Mukta, M. S. H., Akter, A., Ahmed, M., & Islam, S. (2023). A Review on Deep-Learning-Based Cyberbullying Detection. Future Internet, 15(5), 179.
11. Fryer, W.A. (2006, November 20). Addressing cyberbullying in schools. The TechEdge: The Journal of the Texas Computer Education Association. Retrieved March 15, 2009.
12. Mesch, G. S. (2009). Parental mediation, online activities, and cyberbullying. CyberPsychology & Behavior, 12(4), 387-393.
13. Şahin, M. (2012). The relationship between the cyberbullying/cybervictmization and loneliness among adolescents. Children & Youth Services Review, 34(4), 834-837.
14. Jackson, C. (2011). Your students love social media ... and so can you Teaching Tolerance, 39, 38-41.
15. Varjas, K., Talley, J., Meyers, J., Parris, L., Cutts, H., & Hankin, A. (2010). High school students' perceptions of motivations for cyberbullying: an exploratory study. Western Journal of Emergency Medicine, 11(3), 269-273.