A Creativity Survey of Cyberbullying Classification based on Social Network Analysis

Mengran Liu

mengran2023@163.com

School of Engineering, New South Wales University

Abstract: Cyberbullying has been a relatively common problem in recent years. With the development and popularity of the Internet and social media, harassment, discrimination and verbal attacks on the Internet have become more and more frequent, leading to the mental health of many Internet users. In addition, the rise and application of neural networks and text mining have successfully contributed to the research on the detection and prediction of cyberbullying. In this paper, we study the detection methods of cyberbullying and summarize its research status and progress. The research questions, research methods, and measurement methods in the target papers and references were classified. Finally, all the studies are summarized and the directions and innovations of the unstudied are listed.

Keywords- Social network analysis, cyberbullying, neural networks, text mining

1. Introduction

The development of Internet technology has made information dissemination more rapid, but the negative effects of Internet expression have also been criticized due to its anonymity, zero threshold and immediacy. The Internet has amplified certain dark forces in human nature, infinitely contributing to the chaotic situation of online incitement and attacks. In the common space of online social networking, insulting and abusive language, personal attacks, malicious reports and other cyber-bullying phenomena appear from time to time, which deserve attention. The impact on teenagers is particularly serious. Due to their immature mental development, they are more likely to imitate the bad behaviors on the Internet, and bullying teenagers are more likely to cause negative problems in reality, such as mental illness, violent tendencies, alcoholism, etc. This unhealthy online phenomenon can cause a vicious circle in society, so accurate and effective detection of cyberbullying is important to stop more people from being harmed and to create a healthy Internet environment.

In the computer science community, many algorithms and models for detecting cyberbullying incidents have been studied, with the most researched being the training and detection of online speech through natural language processing. The more traditional approach is to achieve bullying detection or classification by first selecting a dataset, then going through data pre-processing to obtain an experimental corpus, followed by feature recognition and model training. Among them, feature recognition mostly uses Term Frequency-Inverse Document Frequency (TF-IDF) technique, while model training utilizes various machine-learning models, such as Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and Long short-term

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334259

memory (LSTM)[4]. However, the fact is that the identification of bullying is much more than that. There is a lot of bullying language that is not very recognizable because it does not necessarily use words with violent connotations directly, but is more metaphorical and more targeted in its innuendo. For example, if a user uses cyber violence against another user, the word "ugly" is easily recognizable if he says "you're ugly," while when he says "you look like a pig, it's really Funny", there is no obvious offensive word here, so it is not easy to be recognized. Therefore, this survey also includes some identification of online violence based on user characteristics and connections, such as building community graphs between users and semantic modeling based on users' usual information and records. In this survey, we present a variety of research on cyberbullying in the computer domain. For each of the techniques discussed, we have provided support through an overview, and the information provided here is -tended to provide an overview and categorical summary of the field.

The rest of the paper is organized as follows. Section II gives the classification of research objects of cyberbullying. Section III introduces the classification of research methods. Section IV introduces the comparison of experimental analysis in related literature. Section V discusses the research opportunities in future work and Section VI concludes the paper.

2. Classification of Research Objects

Table 1: Different Research Objects

User Objectives	Data Feature	
	Without user relationship	**With user relationship**
Cyberbullying detection	I. [1][8][9][10][12]	II. [2][7][11][14][15]
Cyberbullying predictions	III.	IV.

2.1. Criteria

Since the methods can vary greatly when experimented with different research subjects, they can be divided into those that use language as the object of detection and those that use the user as the object of detection. In addition, the algorithm has many implementation purposes, which are mainly divided into bullying detection purposes and prediction purposes. In this section, two independent and different criteria would be used to divide research objects into different types:

1) **Topics**. There are two kinds of purpose of building relationships here: **Detection** or **Predictions**. Detection aims to better analyze the phenomenon of social network bullying, usually by extracting features from the already existing information and then training an effective model to detect the phenomenon of cyberbullying, while prediction is based on detection to be able to predict the occurrence of the bullying phenomenon one-step earlier.

2) **Data type**. There are two types here: **Use user relationship** or **No user relationship**. Traditional detection methods are trained by published content, so most of the extant articles use this approach. However, experimental methods with a user perspective have also emerged,

specifically by analyzing individual user information, language patterns, and user-user interactivity to detect verbal bullying[13].

2.2. The Classification

Based on the appeal classification standard, we give the classification in Table 1. The meaning of each class is as follows:

2.2.1. Without user relationship & Cyberbullying detection

References ([1] [8] [9] [10] [12]) belong to Type I. This type is no bullying detection using relationships between users.

Reference [1] sequentially detected whether the four types of bullying language: curse, insult, sexual and threat, were spoken online by building a corpus, data preprocessing, feature identification (TF-IDF), and model training (SVM) experiments.

Reference [8] aims to detect whether each text belongs to the following three bullying language types: Covertly-Aggressive (CAG), Overtly-Aggressive (OAG), and Non-Aggressive (NAG).

Reference [9] focuses on the aggressive detection of network speech by using deep learning methods. The study provides a troll detection framework by manually extracting features using unigram and bigram, and then feeding feature selection and significant features to the dense layer of Multilayer Perceptron.

Reference [10] analyzed and trained the relationships and differences existing between different kinds of cyberbullying languages, from which a detection system applicable to all cyberbullying languages was concluded.

Reference [12] used gender specificity to improve the accuracy of bullying language differentiation. Using the collected posts of internet users on social networks as a dataset, the insulting language for women and men was preprocessed, feature identified, model trained for the dataset separately, and the final detection accuracy was higher than the detection accuracy regardless of gender.

2.2.2. With user relationship & Cyberbullying detection

References ([2] [7] [11] [14]) belong to Type II. This type is bullying detection using user relationships.

Reference [2] creates a community graph by user features and user relationships, experiments on a dataset of 16,000 tweets based on feature analysis of the community, and then uses a modified Recurrent Neural Network (RNN) model for content detection and identification.

Reference [7] detects bullying by studying the relationships between online users and their interactions, and analyzing the association of online bullied users with offline characteristics.

Reference [11] used features of social network connections to put user-to-user bullying behaviors and proposed an integrated model dealing with users and graph-theoretic features for targeted detection of cyberbullying to predict the dynamics of cyberbullying in social networks.

Reference [14] proposes a principled graph-based approach for modeling temporal dynamics and topic consistency throughout user interactions. Empirically evaluate the effectiveness of the approach through the tasks of session-level bullying detection and comment-level case studies.

Reference [15] subsets tweets based on social network relationships,collects all nodes and generates a graph based on user information files.

2.2.3. With user relationship & Cyberbullying prediction

This type is not using user relationships to predict bullying. No references belong to Type III.

2.2.4. Without user relationship & Cyberbullying prediction

This type is using user relationships to predict bullying. No references belong to Type IV.

3. Classification of Research Methods

Table 2. Different Research Methods

Datasets Feature	Model training method	
	Single model	Hybrid Models
Multiple datasets	I. [5][14]	II. [10][11]
Single datasets	III. [1][7][8][12][15]	IV. [2][9]

3.1. Criteria

For the study of cyberbullying, they used different approaches to experiment with data sets, training models, etc. In terms of datasets, some use a whole dataset, some use different kinds of datasets to compare experiments, while in terms of building models there exist traditional training models as well as models that use a combination of multiple algorithms. In this section, two independent and different criteria would be used to divide research objects into different types:

1) **Datasets**. There are two types here: **Multiple datasets** or **Single datasets**. A single dataset represents that only one dataset was used for each experiment, including the same data source or multiple data sources mixed to form one dataset. Whereas, multiple datasets mentioned in the classification represent experiments in which different datasets were used and the experimental results of each dataset were controlled separately.The dataset is drawn from established datasets and currently popular social media, such as Twitter[3].

2) **Model training method**. There are two kinds of method here: **No portfolio model** or **Existence of portfolio models**. Training models for text are divided into traditional machine learning models and models combining multiple algorithms.

3.2. The Classification

Based on the appeal classification standard, we give the classification in Table 2. The meaning of each class is as follows:

3.2.1. Single model & Multiple datasets

References ([5] [14]) belong to Type I. This type is based on multiple datasets and does not exist portfolio models to complete the training.

Reference [5] proposed a new and more advanced model based on existing models and algorithms: node2vec, a semi-supervised algorithm for scalable feature learning in networks. This model uses a network constructed with edges and nodes for edge prediction and labeled multi-classification tasks, validated and evaluated in experiments using a variety of datasets.

Reference [14] presents a principled graph-based approach for modeling temporal dynamics and topic consistency throughout user interactions. Two datasets, Instagram and Vine, are used for bullying detection separately, and then the experimental results of the two datasets are compared.

3.2.2. Hybrid models & Multiple datasets

References ([10] [11]) belong to Type II. This type is based on multiple datasets and exists portfolio models to complete the training.

Reference [10] uses eight datasets of different bullying categories from Twitter, Formspring, Kaggle, the forum Stormfront, and Wikipedia. First generalization experiments were conducted on these datasets, and the models used two bilayers LSTMs. however, the results trained on each dataset appeared to vary significantly, highlighting the unreliability and unpredictability of transferring pre-trained cyberbullying classifiers to new communities is unreliable and unpredictable. Finally, the implemented integrated models were again combined into an accurate and general classifier, and the datasets were merged into one and retrained. The experiments conclude that the conflicting ways in which many neutral words are labeled between datasets make it difficult to migrate the system, which would be solved if the system were allowed to train on as diverse a dataset as possible.

Reference [11] experiments using two datasets as well as comparisons, the Myspace dataset and Formspring dataset both used a C4.5 decision tree classifier for training bullying detection. Finally, an aggregate model for global cyberbullying dynamics is proposed for further study of global cyberbullying dynamics.

3.2.3. Single model & Multiple datasets

References ([1] [7] [8] [12][15]) belong to Type III. This type is based on single datasets and does not exist portfolio models to complete the training.

Reference [1] crawled Facebook page information as a dataset and trained it using SVM model.

Reference [7] defines several hypotheses linking social characteristics to cyberbullying. These hypotheses were tested through user relationship graphs in the Twitter CAW 2.0 corpus, which were then manually labeled, and the analysis revealed that most of the identified hypotheses apply to very different user settings.

Reference [8] crawled a dataset from Facebook and experimented sequentially using multiple training models including plain Bayes, decision trees, SVM, MLP, LSTM and CNN, with CNN working the best.

Reference [12] used MySpace as a dataset and used a support vector machine to classify different genders for bullying language selection and used gender specificity to improve the accuracy of bullying language differentiation.

Reference [15] obtained datasets from Twitter and experimented with several training models in turn, including Random Forest, SVM, Logistic Regression, AdaBoosting, Parsimonious Bayes, Stochastic Gradient Descent (SGD), CNN and LSTM.

3.2.4. Hybrid models & Single datasets

References ([2] [9]) belong to Type IV. This type is based on single datasets and exists portfolio models to complete the training.

Reference [2] used a dataset from Waseem and Hovy, already collated in 2016, to create community graphs with nodes for authors, and author profiles using the node2vec framework, and then classify the content into racist vs. sexist (modified RNN with softmax instead of sigmoid).

Reference [9] used a combination of CNN-BiLSTM and CNN-LSTM to detect attacks based on the cyber-trells dataset, and then implemented bullying detection through data preprocessing, feature extraction TF-IDF, feature selection, classification, training, and evaluation.

4. Review of Experimental Analysis

In this section, we will classify the metric of evaluation and system factors, as shown in Table 3. In Table 3, all experimental analysis is also classified according to the metric and factors. It can be seen from Table 3 that most of the references compare accuracy, recall rate, f-measure, precision, and others.

Table 3. Experiments with Different Metric and Factors

Metric	System Factors			
	Algorithm	Datasets	Tag	Feature
Accuracy	[8][9][15]		[1][8]	[11][12]
Recall Rate	[2][8][9][14]	[10][14]	[1][2][8]	[12]
F-measure	[2][5][8][9]	[5][10]	[1][2][8]	[12]
Precision	[9][2]		[1][2]	[12]
Others	[5][14]	[5][14]		

4.1. Metric of Evaluation

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. The formula is as follows[6]:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Recall Rate means the proportion of actual positives was identified correctly. The formula is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F-Measure is a statistical quantity, F-Measure is also known as F-Score, F-Measure is a weighted summation average of Precision and Recall, which is a common evaluation criterion in the field of IR (information retrieval) and is often used to evaluate the goodness of classification models. The formula is as follows:

$$\text{F}_\beta = \frac{(\beta^2 + 1)\text{PR}}{\beta^2 \cdot \text{P} + \text{R}}$$

Where is the parameter, P is the Precision and R is the Recall. When the parameter = 1, it is the most common F1-Measure:

$$\text{F}_1 = \frac{2 \cdot \text{PR}}{\text{P} + \text{R}}$$

Precision means the proportion of correct positive identification. The formula is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Other metric includes UAC and Time.

4.2. System Factors

Algorithms represent instructions containing clearly defined functions for computing. When run, it can start with an initial state and an initial input (possibly empty), go through a finite and well-defined set of states to produce an output and stop at a final state.

Dataset represent an abstract class used to package data.

Tag represents a way of organizing Internet content, and is a highly relevant keyword that helps people easily describe and categorize content. In this article specifically represents for the naming of bullying categories.

Feature represents a custom classification feature method.

4.3. Experimental Comparison

In reference [1], the author compares accuracy, recall rate, precision, and F-measure under different tags. The experimental dataset is used 70% for training the SVM model and 30% for evaluation, using tenfold cross-validation, with recall and precision rates staying up throughout

1 to 10 iterations, reaching 0.88 and 0.8 respectively. In addition, the classification results for more than half of the bullying language types were insult, followed by sexual and threat, with the curse being the least.

In reference [2], the author compares precision, recall rate, and F-measure under different tags and algorithms. The results of these three metrics also show intuitively that the authors' algorithm has better data results compared with the established algorithm, except for the precision in the Racism class.

In reference [5], the author F-measure and time under different datasets and algorithms.

In reference [8], the author compares the F-measure, accuracy, and recall rate under different datasets, tags, and algorithms. The CNN model has the highest accuracy and recall with 0.73 and 0.59, respectively, and the SVM Model with L2 penalty has the highest precision with 0.59. In addition, for the F1-score, both the LSTM and CNN models have the same 0.58.

In reference [9], the author compares the F-measure, accuracy, precision, and recall rate under different algorithms. The model MLP with TF-IDF outperformed the other models with 92% accuracy. On the other hand, MLP using word embedding and CNN-BiLSTM has an accuracy of 87%, which is higher than the accuracy of 86% for CNN-LSTM. CNN-LSTM and CNN-BiLSTM show recall values of 78% and 83%, respectively, which are lower than the proposed model with 90% recall, and show an accuracy value of 93%, which is higher than the proposed 90% accuracy method. Since the F1 score is the balance of precision and recall values, the proposed method in the article achieves the highest value with a 90% F1 score.

In reference [10], the authors compare the F-measure and recall rate under different datasets and analyze them separately by integrating the experimental results of the experimental and generalized models.

In reference [11], the author compares the accuracy under different features, and the detection results are consistent on both datasets.

In reference [12], the author compares accuracy, recall rate, F-measure, and precision under different gender specificity. Since women's language is more euphemistic, the algorithm is more accurate for male-specific experimental results.

In reference [14], the author compares the recall rate and AUC under different algorithms and datasets.

In reference [15], the authors compare the accuracy of the training results of different models and the final result is that SVM has the best performance.

There are no references to measure accuracy and prediction in different data sets. This is because such comparisons between different datasets are meaningless. The data in the real application are inherently uncertain, so the experimental accuracy and prediction is used as evaluation metrics for the experimental model, not for the datasets.

5. Discussion and Suggestion

This paper discusses the research methods and research objects of various references and finds that there are still some issues about the treatment of cyberbullying that have not received the attention of researchers. Therefore, this paper puts forward the following directions, which can provide directions for future text mining research:

1) Predicting cyberbullying based on user relationships.

By analyzing and sub-predicting user information and network relationships between users in social media to predict in advance whether they will commit cyber violence against other users in the future. For example, if it is detected that a user has frequently been abusive to a specific category of users in the past, then this category of users will be classified into the predicted list.

2) Text mining based web bullying prediction.

The training and analysis of users' speech in social media is used to predict in advance whether they will commit cyber violence against other users in the future time period. For example, it detects the linguistic tendencies of users during their interaction in chat or comments, and thus predicts the linguistic trends of their interactions.

3) Different metric and factors.

The feasibility of the bullying detection method can be assessed by comparing other metrics (accuracy, precision, recall, and F-measure) under different Tag or Features.

6. Conclusions

A previous analysis will show that most studies on cyberbullying go to detect existing bullying language, and even though a part of the studies also mention 'prediction', in reality, their work does not give predictions before the speech is made, but only detects new occurrences of bullying through trained models. Alternatively, in [11] it was envisioned to predict bullying through user relationships, but ultimately it did not work. So that there is little analysis and research in this area of cyberbullying prediction. Therefore either by text detection or user relationship approach, cyberbullying prediction is meaningful research.

References

[1] K. D. Gorro, M. J. G. Sabellano, K. Gorro, C. Maderazo, and K. Capao,"Classification of cyberbullying in facebook using selenium and SVM," in 2018 3rd International Conference on Computer and Communication Systems (ICCCS), pp. 183–186, IEEE, 2018.

[2] P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova, "Author profiling for abuse detection," in Proceedings of the 27th International Conference on Computational Linguistics, (Santa Fe, New Mexico, USA), pp. 1088–1098, Association for Computational Linguistics, Aug. 2018.

[3] https://developer.twitter.com/en/docs.

[4] M. Stamp, A Survey of Machine Learning Algorithms and Their Application in Information Security: An Artificial Intelligence Approach, pp. 33–55. 09 2018.

[5] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” vol. 2016, pp. 855–864, 07 2016.
[6] https://developers.google.com/machine-learning/crash-course/classification/accuracy.
[7] Q. Huang, V. Singh, and P. Atrey, “On cyberbullying incidents and underlying online social relationships,” Journal of Computational Social Science, vol. 1, 09 2018.
[8] V. Singh, A. Varshney, S. S. Akhtar, D. Vijay, and M. Shrivastava, ”Aggression detection on social media text using deep neural networks,” in Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 43–50, 2018.
[9] S. Sadiq, A. Mehmood, D. S. Ullah, M. Ahmad, G. S. Choi, and B.-W.On, “Aggression detection through deep neural model on twitter,” Future Generation Computer Systems, vol. 114, 07 2020.
[10] K. Richard and L. Marc-André, “Generalisation of cyberbullying detection,” arXiv preprint arXiv: 2009.01046, 2020.
[11] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, “Identification and characterization of cyberbullying dynamics in an online social network,” in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 280–285, 2015.
[12] M. Dadvar, F. d. Jong, R. Ordelman, and D. Trieschnigg, “Improved cyberbullying detection using gender information,” in Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), University of Ghent, 2012.
[13] Q. Huang, V. Singh, and P. Atrey, “On cyberbullying incidents and underlying online social relationships,” Journal of Computational Social Science, vol. 1, 09 2018.
[14] S. Ge, L. Cheng, and H. Liu, “Improving cyberbullying detection with user interaction,” in Proceedings of the Web Conference 2021, pp. 496–506, 2021.
[15] A. Wang and K. Potika, "Cyberbullying Classification based on Social Network Analysis," 2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService), 2021, pp. 87-95, doi: 10.1109/BigDataService52369.2021.00016.