How to identify Cyberbullying with Machine Learning

M. L. Fujimoto, M. Gaseta, S. O. Rezende, R. A. F. Romero

Universidade de São Paulo, Brazil
mlika@usp.br, mgaseta@usp.br, solange@icmc.usp.br, rafrance@icmc.usp.br

Abstract. Cyberbullying is a form of bullying that has emerged and is a concerning problem with the exponential increase of social media users. Social networks provide a suitable environment for those bullies to attack and cause serious psychological problems in their victims. To mitigate these issues, proactive measures are essential to detect and prevent cyberbullying before disseminating harmful content. With this concern in mind, this article proposes an approach to combine TF-IDF with machine learning models to automatically identify cyberbullying. These models are evaluated using metrics such as accuracy and F1-score to identify and classify cyberbullying instances. The research aims to contribute to the development of automated systems capable of preemptively addressing cyberbullying on social media platforms.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: machine learning, cyberbullying, natural language processing

1. INTRODUCTION

In recent years, the proliferation of social media platforms has revolutionized the way people communicate and share information. With millions of active users, social media platforms serve as a significant channel for public discourse, breaking news, and social interaction. However, this openness and immediacy have also led to a rise in offensive content, ranging from hate speech and bullying to more subtle forms of harassment and discrimination [Shome and Kar 2021].

Offensive posts are not merely harmful to the individuals directly targeted, they contribute to a broader societal impact. Such content can perpetuate stereotypes, incite violence, and create a hostile online environment that discourages free and open dialogue. The psychological impact on victims can be profound, leading to stress, anxiety, and in severe cases, self-harm or suicide [Balayn et al. 2021]. Moreover, the pervasive nature of offensive content can undermine the integrity of online communities, reducing trust and engagement among users [Vasalou et al. 2008].

Given the scale and rapid pace of content generation, manually moderating offensive posts is both impractical and insufficient. Consequently, there is a pressing need for automated solutions that can efficiently and accurately identify and classify offensive content [Kumar and Sachdeva 2019]. Machine learning techniques offer a promising approach to address this challenge. By leveraging large datasets and sophisticated algorithms, machine learning models can learn to detect nuanced patterns in text, enabling the classification of posts into various categories of offensiveness.

This study explores the application of several machine learning methods to classify posts into different offensive classes. Specifically, we employed Support Vector Classification, Multinomial Naive Bayes, Multi-layer Perceptron Classifier, Decision Tree Classifier, and Random Forest Classifier, all implemented using the scikit-learn library. To compare the performance of these models, we used standard metrics like Accuracy and F1.

The remainder of this article is organized as follows: In Section 2, some related works on the detection of offensive content in social media are described. In Section 3, the dataset, pre-processing steps, and the machine learning models used are presented. The experimental setup, results, and the analysis performed are presented in Section 4. Finally, in Section 5, some findings are summarized and presented directions for future research.

2. RELATED WORK

The field has garnered significant attention due to the rising prevalence of harmful online behavior and its negative impact on individuals and society. We present a selection of key studies that have contributed to the advancement of techniques and methodologies for detecting offensive content in social media.

Davidson et al. [2017] conducted a seminal study in which they developed a dataset of tweets labeled as hate speech, offensive language, or neither. They employed a logistic regression classifier to distinguish between these categories, achieving an accuracy of 91%. Their work highlighted the importance of feature engineering, particularly the use of n-grams and part-of-speech tags, in improving classification performance.

Waseem and Hovy [2016] provided a comprehensive analysis of hate speech detection by creating a dataset annotated with various forms of offensive content, including racism and sexism. They explored the use of different classifiers such as Support Vector Machines (SVM), achieving an accuracy of 73.89%. They demonstrated that incorporating domain-specific knowledge, such as the gender and ethnicity of the tweet author, could enhance detection accuracy.

Badjatiya et al. [2017] investigated deep learning approaches for hate speech detection on Twitter using Waseem and Hovy [2016] dataset. They compared traditional machine learning models with deep learning techniques, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. Their findings indicated that deep learning models, particularly CNNs, outperformed traditional methods with an accuracy of 93.88%, particularly when leveraging word embeddings for feature representation.

Zhang et al. [2018] focused on the problem of offensive language detection by proposing a hybrid approach that combined deep learning and traditional machine learning techniques. They utilized a bidirectional LSTM network for feature extraction and a gradient boosting classifier for final classification, achieving an accuracy of 95.25%. Their approach demonstrated improved performance over standalone models, emphasizing the potential of hybrid methods in this domain.

Founta et al. [2018] created a large-scale dataset of tweets annotated with multiple types of abusive language, including hate speech, aggression, and cyberbullying. They evaluated several machine learning algorithms, including Random Forests and SVMs, with the best-performing model achieving an accuracy of 82%. They stressed the importance of having diverse and balanced datasets to train robust classifiers. Their work underscored the complexity of offensive content and the need for multi-faceted approaches to effectively address it.

There are other types of studies, such as those focused on mobile [Thun et al. 2022], focused on specific languages like Arabic [Haidar et al. 2019], or hybrid [Almomani et al. 2024]. Many works focus on deep learning, such as Kompally et al. [2021] or Dadvar and Eckert [2020], but despite their excellent performance, models based on deep learning depend on a huge amount of annotated data and high computational power to train those models.

While the previous studies mentioned above have significantly advanced the field of offensive content detection, our study contributes further by implementing and comparing multiple models, including Support Vector Classification, Multinomial Naive Bayes, Multi-layer Perceptron Classifier, Decision Tree Classifier, Random Forest Classifier. Unlike prior studies that often focused on a single or a few methods, our comprehensive approach evaluates a broader range of classifiers considering binary and multi-class classification. We specifically use accuracy as the primary evaluation metric to ensure a consistent and comparative analysis of model performance. Our contribution also includes an extensive evaluation of the models' effectiveness in classifying various offensive classes, providing deeper insights into their relative strengths and weaknesses in the context of social media content moderation.

3. MATERIALS, METHODS AND PROPOSED APPROACH

This section details the dataset, pre-processing steps, and the machine learning models used in our study. It is divided into the following subsections: 3.1 Presenting the Data, 3.2 Exploration and Pre-processing, 3.3 Classification Models, and 3.4 Proposed Approach.

3.1 Presenting the Data

In light of the challenges associated with cyberbullying detection, the dataset containing over 47,000 tweets labeled according to the class of cyberbullying was presented [Wang et al. 2020]. This dataset was mainly chosen for its focus on cyberbullying and for having classes directly related to the topic. Other datasets, such as [Salawu et al. 2021], have different labels like Bullying, Insult, Profanity, Sarcasm, and others, as they focus on mixed areas such as cyberbullying and abuse detection. The classes, along with the number of entries for each one, are shown in Figure 1.

The dataset is available in English and is divided into two columns: the text information without treatment (tweet) and the class it belongs to. Figure 2 illustrates an example of tweets and their corresponding classes. Despite efforts to balance the data, the predominance of cyberbullying content in most tweets, especially within the mixed-content classes like "Other Cyberbullying" and "Not Cyberbullying", may significantly impact the model's performance.

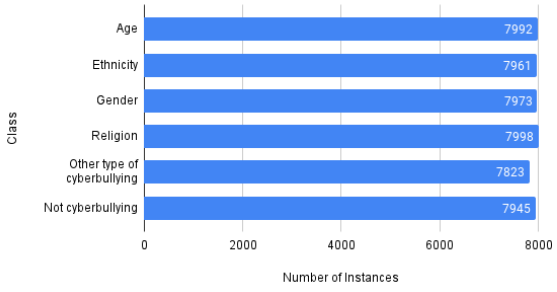

Fig. 1. Classes and number of instances.

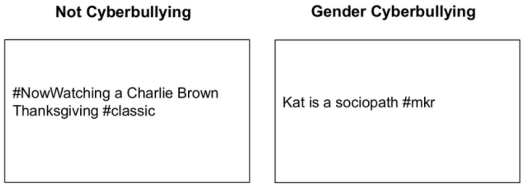

Fig. 2. Example of tweets and their classes.

3.2 Exploration and Pre-processing

Upon exploration, no null entries were found in the dataset, ensuring the integrity of the data. The class names were converted to numerical values as shown in Table I.

Class Label	Numerical Value
religion	1
age	2
gender	3
ethnicity	4
not_cyberbullying	5
other_cyberbullying	6

Table I. Mapping text codes to numbers.

Additionally, a new column was added for binary classification to indicate whether the entry is cyberbullying (1) or not (0). It is important to note that this binary class is imbalanced, with a majority of entries likely falling under the cyberbullying category (1). Therefore, careful pre-processing

and training procedures are essential to mitigate these challenges effectively. Pre-processing steps included normalizing the text data of the tweets:

—Lowercasing every character: To ensure uniformity and reduce the complexity of text processing;
—Removing tags, URLs, punctuation, and special characters: To eliminate irrelevant content that does not contribute to the semantic meaning;
—Removing emojis: Emojis can play an important role in the context of sentiment analysis, but in this work we will focus on offensive texts and, for this reason, they were considered non-essential;
—Removing stopwords: Common words like 'and', 'the', etc., which do not add significant meaning in this context and can dilute the focus of the analysis;
—Lemmatizing: Converting words to their base or dictionary form to standardize variations of words.

Term Frequency-Inverse Document Frequency (TF-IDF) was used to vectorize the text data for training and testing. TF-IDF was choosen because of its simplicity and speed. It helps in converting text data into numerical vectors while emphasizing the importance of less frequent but significant words. TF-IDF assigns weights to words, making it a straightforward method for document representation in information retrieval and machine learning tasks [Samatha et al. 2023].

As mentioned above, the new column for binary classification is imbalanced. Imbalance is a common scenario in many datasets. Such scenarios pose challenges for accurate classification by machine learning models. To address the potential imbalance in the dataset, two experiments were conducted: one involving the application of the Synthetic Minority Over-sampling Technique (SMOTE) treatment [Chawla et al. 2002] with Stratified K-Fold cross-validation, and another approach considering only Stratified K-Fold cross-validation.

SMOTE generates synthetic samples for the minority classes, which helps in creating a balanced training dataset. This technique was chosen because it has been successfully applied in different domains. Another technique, Stratified K-Fold cross-validation is used to validate machine learning models, especially in the context of imbalanced datasets. The dataset is divided into K subsets (folds) of approximately equal size. Each fold is used once as a validation set while the remaining K-1 folds are used for training. Stratified K-Fold maintains the percentage of samples for each class in every fold to ensure that each fold is representative of the overall class distribution.

3.3 Classification Models

In this subsection, the classification models used in our study are briefly presented with a small description of why they were chosen and their hyperparameters (for the hyperparameters not specified, the default values were used). The models were implemented using Scikit-Learn, and the names of the libraries and functions used will also be mentioned.

Support Vector Classification (SVC from `sklearn.svm`): SVC aims to find the hyperplane that best separates the classes in the feature space. A linear kernel was chosen for its simplicity and effectiveness in text classification tasks.

Multinomial Naive Bayes (MultinomialNB from `sklearn.naive_bayes`): This classifier is based on Bayes' theorem and is particularly effective for text classification, benefiting from its simplicity and effectiveness in handling discrete data.

Multi-layer Perceptron Classifier (MLPClassifier from `sklearn.neural_network`): An MLP is a type of neural network that consists of multiple layers of nodes. It is capable of learning complex patterns in the data, making it suitable for tasks requiring non-linear decision boundaries.

Decision Tree Classifier (DecisionTreeClassifier from `sklearn.tree`): This model splits the data into subsets based on the feature that provides the highest information gain, making decisions that

lead to the most homogeneous subgroups.

Random Forest Classifier (RandomForestClassifier from `sklearn.ensemble`): A Random Forest is an ensemble method that constructs multiple decision trees during training and outputs the class that is the mode of the classes predicted by individual trees. It is robust to overfitting and handles large datasets well.

It is important to highlight that there are variations of these algorithms beyond those used in this work. Some of these variations can be configured by adjusting the models' hyperparameters.

3.4 Proposed Approach

An overview of the approach can be seen at Figure 3. The first step was pre-processing, after that we conducted experiments considering both binary and multi-class classes. Strategies for imbalance were also considered and a variety of models were evaluated.

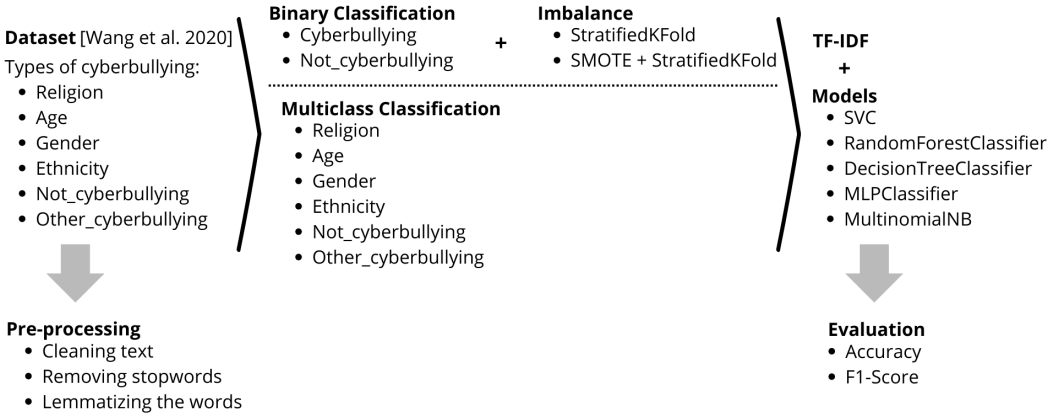

Fig. 3. Overview of the methodology.

All tasks and evaluations were designed to run on a computer with limited access to a GPU. To evaluate the model's robustness while minimizing computational cost, we employed 3-fold cross-validation. This size of k-fold was chosen considering the restrictive environment available for this experiment. The performance of the models was assessed using two metrics: F1-score and accuracy. Accuracy measures the overall correctness of the model, while the F1-score provides a balance between precision and recall, making it suitable for imbalanced datasets.

4. EXPERIMENTS

In this section, the experimental setup and results for both binary and multi-class classification of cyberbullying tweets are presented and discussed[1]. The experiments aim to evaluate the performance of various machine learning models in detecting offensive content on social media platforms.

The hypothesis is that using TF-IDF for feature extraction in the context of cyberbullying detection, combined with machine learning models like Random Forest Classifier, can classify cyberbullying with high accuracy even with few resources. The motivation for using TF-IDF is because of the simplicity and effectiveness in extracting relevant features from text. This is crucial in domains like cyberbullying detection where labeled data may be difficult to obtain.

[1]Code available at https://github.com/ds2024m/cyberbullying.

In every experiment described below, TF-IDF and a classifier were employed. For binary classification, two experiments were conducted: one using SMOTE (to handle imbalances) combined with StratifiedKFold, and another using only StratifiedKFold. In this experiment, we aim to analyze the impact of using balanced synthetic data, by SMOTE technique, in terms of its performance compared with the performance obtained by StratifiedKFold. The results of the experiment with SMOTE can be seen in Table II. Meanwhile, the experiment with StratifiedKFold showed better performance in terms of F1 and accuracy and is detailed further in Table III.

In Table II, it can be observed that all models showed results ranging from 0.81 to 0.83 for Accuracy and from 0.88 to 0.89 for F1-Score. Specifically, the Random Forest model achieved the highest accuracy of 0.83 and an F1-Score of 0.89, indicating strong generalization capability and performance in cyberbullying detection with the use of SMOTE for class balancing.

Model Name	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVC	0.82 ± 0.0021	0.96	0.82	0.88
MultinomialNB	0.81 ± 0.0041	0.94	0.83	0.88
MLPClassifier	0.82 ± 0.0015	0.90	0.88	0.89
Decision Tree	0.82 ± 0.0020	0.91	0.86	0.89
Random Forest	0.83 ± 0.0029	0.92	0.86	0.89

Table II. Binary classification results using SMOTE and StratifiedKFold.

In Table III, using only StratifiedKFold, the results were slightly different, ranging from 0.82 to 0.87 for accuracy and from 0.90 to 0.92 for F1-Score. Despite some differences, all models maintained robust performances, indicating a good balance between precision and recall in the classification.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVC	0.87 ± 0.0026	0.81	0.68	0.92
MultinomialNB	0.85 ± 0.0018	0.79	0.63	0.92
MLPClassifier	0.82 ± 0.0011	0.70	0.69	0.90
Decision Tree	0.84 ± 0.0026	0.70	0.69	0.90
Random Forest	0.86 ± 0.0014	0.74	0.68	0.91

Table III. Binary classification results using only StratifiedKFold.

An experiment was conducted considering the original classes. For multi-class classification, the scalability and computational efficiency of the classifiers may become more critical as the number of classes increases. Additionally, the performance of the classifiers may be affected by the presence of class imbalances within the multi-class dataset, potentially impacting the precision and recall for individual classes. The choice of classifier and its ability to handle multi-class scenarios can significantly impact the overall predictive accuracy.

Table IV shows that SVC exhibited the best performance among the evaluated classifiers. As mentioned earlier, the mixing of classes can significantly influence a classifier's performance. This can be observed because of the performance in classes 5 and 6, "Not Cyberbullying" and "Other Cyberbullying", indicating the struggle of the models with those types of classes and it impacts the overall evaluation of the classifier.

The experiments showed that StratifiedKFold achieved robust results in dealing with imbalanced data, even without the combined use with SMOTE. The explanation for this result lies in the fact that SMOTE generates synthetic data that may not accurately represent the true data distribution encountered by the model in a production environment. This discrepancy can lead to performance that falls below expectations, as demonstrated by these results.

The results also confirmed our hypothesis that it is possible to use TF-IDF combined with a machine learning model to achieve high accuracy even with limited computational resources. These results

Model	Class	Precision	Recall	F1-Score	F1 Weighted Avg	Accuracy
SVC	Religion	0.96	0.96	0.96	0.85	0.85
	Age	0.96	0.98	0.97		
	Gender	0.90	0.87	0.89		
	Ethnicity	0.97	0.98	0.98		
	Not Cyberbullying	0.62	0.57	0.59		
	Other Cyberbullying	0.65	0.70	0.68		
MultinomialNB	Religion	0.82	0.97	0.88	0.74	0.77
	Age	0.71	0.98	0.82		
	Gender	0.81	0.86	0.84		
	Ethnicity	0.83	0.93	0.88		
	Not Cyberbullying	0.66	0.39	0.49		
	Other Cyberbullying	0.69	0.45	0.54		
MLPClassifier	Religion	0.95	0.94	0.95	0.80	0.80
	Age	0.95	0.95	0.95		
	Gender	0.84	0.85	0.85		
	Ethnicity	0.98	0.98	0.98		
	Not Cyberbullying	0.51	0.53	0.52		
	Other Cyberbullying	0.58	0.56	0.57		
Decision Tree	Religion	0.96	0.93	0.95	0.81	0.81
	Age	0.98	0.97	0.98		
	Gender	0.88	0.84	0.86		
	Ethnicity	0.98	0.98	0.98		
	Not Cyberbullying	0.50	0.54	0.52		
	Other Cyberbullying	0.55	0.56	0.55		
Random Forest	Religion	0.96	0.96	0.96	0.83	0.83
	Age	0.98	0.98	0.98		
	Gender	0.91	0.85	0.88		
	Ethnicity	0.99	0.99	0.99		
	Not Cyberbullying	0.59	0.52	0.55		
	Other Cyberbullying	0.58	0.69	0.63		

Table IV. Multi-class classification results.

also highlight the importance of model selection and feature engineering to improve the accuracy and reliability of cyberbullying detection systems, which is crucial for safer online environments.

5. FINAL CONSIDERATION

In this study, we aimed to address the challenge of identifying and categorizing cyberbullying instances effectively with low resources. To achieve that goal, we combined TF-IDF with machine learning approaches for the classification of cyberbullying content on social media platforms. Our experiments highlighted the difficulty of working with imbalanced data, demonstrating that StratifiedKFold was better suited to the scenario than SMOTE, with better performance.

Classifiers such as SVC and RandomForestClassifier exhibited robust performance across binary and multi-class classification tasks, achieving high accuracy and F1-scores. Therefore, this study contributes by presenting experiments with models that require low computational resources, but have good performance results even with limited labeled data and low computational power. It is important to highlight the difficulty in finding labeled data in this context and, therefore, the need for alternatives to large language models that require a huge amount of resources.

For future research, we could explore ensemble methods, other techniques such as word embeddings, lightweight deep learning architectures or adjust pre-trained LLMs to enhance classification accuracy and scalability in large-scale social media datasets. Further research could involve a deeper comparison with related works, including comparative tables of results and computational resources. Furthermore, the research could be expanded to include data from various social media platforms, enabling a more comprehensive evaluation. Addressing the evolving nature of online behavior and linguistic patterns

remains crucial for developing robust and adaptive cyberbullying detection systems. Ultimately, this study contributes to the ongoing efforts in leveraging machine learning for social good, aiming to foster safer and more inclusive online environments.

Acknowledgments: The authors of this paper thank FAPESP (Process 2019 / 25010-5), the National Center for Scientific and Technological Development (CNPq) (process 309575/2021-4) and CAPES.

REFERENCES

Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M. A., Yaseen, Q., and Gupta, B. B. Image cyberbullying detection and recognition using transfer deep machine learning. *International Journal of Cognitive Computing in Engineering* vol. 5, pp. 14 – 26, 2024.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*. pp. 759–760, 2017.

Balayn, A., Yang, J., Szlavik, Z., and Bozzon, A. Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *Trans. Soc. Comput.* 4 (3), 2021.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16 (1): 321–357, 2002.

Dadvar, M. and Eckert, K. Cyberbullying detection in social networks using deep learning based models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 12393 LNCS, pp. 245 – 255, 2020.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*. Vol. 11. pp. 512–515, 2017.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*. Vol. 12, 2018.

Haidar, B., Chamoun, M., and Serhrouchni, A. Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning. *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2019.

Kompally, P., Sethuraman, S. C., Walczak, S., Johnson, S., and Cruz, M. V. Malang: A decentralized deep learning approach for detecting abusive textual content. *Applied Sciences (Switzerland)* 11 (18), 2021.

Kumar, A. and Sachdeva, N. Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. *Multimedia Tools and Applications* 78 (17): 23973–24010, 2019.

Salawu, S., Lumsden, J., and He, Y. A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem (Eds.). Association for Computational Linguistics, Online, pp. 146–156, 2021.

Samatha, B., Karyemsetty, N., Kumar, D. S., Rao, D. K., Mani, G., and Syamsundararao, T. Analysis of a multichannel learning mechanism for speech detection in social networks. *Proceedings of the International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE*, 2023.

Shome, D. and Kar, T. Conoffense: Multi-modal multitask contrastive learning for offensive content identification. In *2021 IEEE International Conference on Big Data (Big Data)*. pp. 4524–4529, 2021.

Thun, L. J., Teh, P. L., and Cheng, C.-B. Cyberaid: Are your children safe from cyberbullying? *Journal of King Saud University - Computer and Information Sciences* 34 (7): 4099 – 4108, 2022.

Vasalou, A., Hopfensitz, A., and Pitt, J. V. In praise of forgiveness: Ways for repairing trust breakdowns in one-off online interactions. *International Journal of Human-Computer Studies* 66 (6): 466–480, 2008.

Wang, J., Fu, K., and Lu, C.-T. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 1699–1708, 2020.

Waseem, Z. and Hovy, D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. pp. 88–93, 2016.

Zhang, Z., Robinson, D., and Tepper, J. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, pp. 745–760, 2018.