Using Sarcasm to Improve Cyberbullying Detection

Xiaoyu Guo, Susan Gauch
University of Arkansas, University of Arkansas
Fayetteville Arkansas, Fayetteville Arkansas
{xsguo, sgauch}@abc.org

Abstract

Cyberbullying has become more prevalent over time, especially towards minority groups, and online human moderators cannot detect cyberbullying content efficiently. Prior work has addressed this problem by detecting cyberbullying with deep learning approaches. In this project, we compare several BERT-based benchmark methods for cyberbullying detection and do a failure analysis to see where the model fails to correctly identify cyberbullying. We find that many falsely classified texts are sarcastic, so we propose a method to mitigate the false classifications by incorporating neural network-based sarcasm detection. We define a simple multilayer perceptron (MLP) that incorporates sarcasm detection in the final cyberbully classifications and demonstrate improvement over benchmark methods.

Keywords: Natural language processing, Machine learning, Cyberbullying detection, Sarcasm detection

1. Introduction

Ever since the increasing popularity of the Internet, people have taken social media as a central place for expressing their opinions, peer reviews, dissemination of scientific information, online discussions and more (Goel and Gupta, 2020). Because of the nature of anonymity in social media, people are more likely to express their own opinions, which do not always agree with other people's opinions. The disagreements can lead to heated discussions, then to hostile arguments. Such arguments can turn into personal attacks, which can ultimately result in cyberbullying as an attempt to perform ad hominem. Cyberbullying is defined as 'an aggressive act or behavior that is carried out using electronic means by a group or an individual repeatedly and over time against a victim who cannot easily defend him or herself (Smith et al., 2008). This behavior can adversely affect a person's mental health, which can lead to social anxiety, depression, stress, and social isolation. Study has shown that people in minority groups are more vulnerable to cyberbullying attack (Llorent et al., 2016), and people with different cultural background may perceive textual context differently, which can cause more confusion and personal attack as the argument goes on.

Many architectures have been proposed to identify and mitigate cyberbullying. Early methods include handmade rules (Bayzick et al., 2011), which achieved an accuracy of 58.63%. Later machine learning based approaches were proposed, including logistic regression(Chavan and Shylaja, 2015) and random forest(Al-Garadi et al., 2016).

More recently, machine learning-based approaches were also proposed, including SVM (Dadvar et al., 2013; Nahar et al., 2013; Zhao et al., 2016) and BERT-based classifiers like HateBERT(Caselli et al., 2020) and CyberBERT (Paul and Saha, 2022). BERT-based classifiers are showing promising results, because they excel in bidirectional textual structure and context, meaning that it takes into account both context to the left and the right when making predictions. The vanilla BERT has been trained on a large corpus, while both HateBERT and CyberBERT have been fine-tuned with cyberbullying datasets.

In this work, we compare several BERT-based benchmark methods for cyberbullying and conduct a failure analysis. We then identify the common characteristics of mis-classified data points to be the use of sarcasm, when the text itself appeared innocent but had a negative intention, or when the text itself appeared hostile but had a positive intention. We address this failure with a sarcasm classifier. Finally, we train a simple multilayer perceptron (MLP) neural network that takes sarcasm into account when classifying cyberbullying, and we demonstrate an improvement in both accuracy and F-1 score.

The remaining part of the paper is organized as follows. In section 2, we highlight existing works on cyberbully detection and sarcasm detection. Then we perform a comparison analysis in section 3. We provide our proposed method and analyze the results in section 4. Lastly, we conclude our paper and identify limitations.

2. Related Work

2.1. Cyberbully Detection

Mahmud et al. (Mahmud et al., 2008)were the first authors that tried to automatically determine cyberbullying text. They constructed a set of rules to extract semantic information used to separate abusive language. Later, Serra and Venter used a neural network to interpret a set of rules that links phone usage patterns among children to cyberbullying activities (Serra and Venter, 2011). Bretschneider et al. included additional profanity features to determine more personalized abusive content, as they believe that such content are more indicative of cyberbullying activities than specific abusive terms (Bretschneider et al., 2014).

Some researchers ventured into the realm of machine learning for automatic cyberbully detection. In 2011, Reynolds et al. used a C4.5 decision tree learner and an instance-based learner to detect language patterns and develop rules to detect cyberbullying content (Reynolds et al., 2011). Stochastic Gradient Descent (SGD) was also used by Al-Garadi et al. to build a cyberbully prediction model (Al-Garadi et al., 2016). Other machine learning techinques used include multinomial Naive Bayes (Stauffer et al., 2012; Hinduja and Patchin, 2008) and Random Forest (Zhao et al., 2016; Lenhart et al., 2010).

Deep learning approaches are also explored. Murshed et al. proposed a RNN-based model with an optimized Dolphin Echolocation Algorithm that fine-tunes RNN's parameters and reduces training time (Chandrasekaran et al., 2022). Roy and Mali developed a transfer learning-based model to prevent image-based cyberbullying issues on social platforms (Roy and Mali, 2022). Fati et al. utilize convoutional LSTM for cyberbullying detection on Twitter (Fati et al., 2023). Alongside the popularity of deep learning, large language models (LLMs) with zero-shot learning abilities can also be used for cyberbully detection task with fine-tuning. One of the most prominent LLM is GPT-3 (Brown et al., 2020) proposed in 2020. Further study can be done on the how well LLMs can solve cyberbullying and sarcasm detection tasks.

Researchers also focused on content-based approaches. Dinakar et al. theorized that clustering the texts by themes first will improve the final classification of cyberbully since the classifiers were able to learn features based on cluster themes like racism, culture, sexuality, and intelligence (Dinakar et al., 2011). Dadvar et al. adopted a similar approach by clustering by writers' gender (Dadvar et al., 2012a). Furthermore, Dadvar hypothesized that incorporating the receiver's action can improve the overall performance(Dadvar et al., 2012b). Such actions include victims replying to the cyberbully post or changing their status on Facebook after receiving a cyberbully text, which can be used to determine the victim's emotional state. In 2020, Balakrishnan et al. conducted a project that incorporates psychological features including personalities, sentiment, and emotion to classify each tweet data into four categories: bully, aggresor, spammer, and normal(Balakrishnan et al., 2020). They used Naive Bayes, Random Forest, and J48 for classification, and they observed that incorporating personalities and sentiments improved cyberbullying detection, but incorporating emotions did not improve the classification result.

More recently, BERT-based approaches have gained popularity. Many projects fine-tuned BERT on cyberbullying datasets which resulted in state-of-the-art performance. Some pre-trained models include CyberBERT(Paul and Saha, 2022), HateBERT(Caselli et al., 2020), and BHF(Feng et al., 2022).

2.2. Sarcasm Detection

One challenging NLP task is sarcasm presented in a sentence, which can cause misconception in the context, and the sentence may not convey the surface meaning and needs further interpretation of the hidden expression. Sarcasm is mainly found in real-life conversations and can be conveyed using body language and facial expressions like an eye roll or tone of speech, but sarcasm also thrives on the Internet. Without the body signal, it is hard to tell if a person is being serious, or they are just using irony. A study in the Journal of Language in Social Psychology has suggested that people tend to use sarcasm more frequently online than in face-to-face interactions(Hancock, 2004). Due to the wide use of sarcasm in social media, sarcasm detection has become a small but interesting research topic niche in NLP.

Similar to cyberbully detection, some sarcasm detection model relies on the use of feature extractions and machine learning. Chatterjee et. al. designed four features used with deep learning models to detect sarcastic sentences (Chatterjee et al., 2020). The features are overtness, acceptability, exaggeration, and comparison. Acceptability is defined as how socially acceptable a sentence is based on the number of unacceptable words, and comparison is the similarities between the compared objects in the sentence using Wu-Palmer similarity (Wu and Palmer, 1994) on Word-net. Overtness and acceptability capture the semantic sense of a sentence. Exaggeration and comparison capture the implicit incongruity, which is between the surface sentiment and the implied sentiment. They found that a Random Forest classifier along with the four features achieved the best performance among the models they trained.

CNNs are another popular model for sarcasm detection. Son et. al. developed a Soft Attention-based BiLSTM in conjunction of ConvNet for sarcasm detection (Kumar et al., 2019). Ashok et. al. also used an LSTM-CNN model to predict sarcasm on processed tweets(Ashok et al., 2020).

3. Cyberbully Detection Model Analysis

3.1. Dataset

We use three different datasets to evaluate cyberbullying classification performance. All three datasets are classified into two classes: cyberbully or non-cyberbully. We name these three dataset by its source: Twitter(Wang et al., 2020), YouTube (Dadvar et al., 2014), and a dataset provided by Kaggle [1].

	cyberbully	non-cyberbully
Twitter	7945	38072
YouTube	417	3047
Kaggle	2806	5993

Table 1: Datasets used for evaluation.

For preprocessing, we remove all data points with less or equal to 4 words. Initial investigation has shown that data points with less than 4 words do not possess enough contextual information to be classified. We also remove all hashtag symbols for each hashtag, and all emojis are replaced with the text provided by the Python *emoji* package. For ethical considerations, we also replaced all users mentions with "@USR", and all URLs are replaced with "URL".

It is worth noting the skew in the dataset. Though with various degrees, all three datasets have more non-cyberbullying data entries than cyberbullying data entries. Skewed datasets are common in cyberbullying datasets, which can hinder the performance of logistic regression or decision tree-based models, since these models rely on class separation and feature correlation. They may not find sufficient features of the minor class data points. Skewed datasets can also cause high accuracy but low F1 score, as the model can classify all testing data into the major class data points, which will achieve a high accuracy, but also a high score of false positive or false negative classifications.

To preserve the imbalance in the dataset, when we randomly split the dataset into training data and testing data, we would first separate each dataset into two datasets, one containing all cyberbullying data and the other containing all non-cyberbullying data. We would randomly select training and testing data from the two sub-datasets, then combine them to form complete training and testing datasets while preserving the distribution of the original dataset.

3.2. Models

First, we want to test pre-existing cyberbully detection models. We choose three different models: the vanilla BERT model, HateBERT, and CyberBERT. We randomly choose 30% of each class to be the testing data, and the remaining 70% will be the training dataset. We fine-tune each model with the training data, and then test the fine-tuned model with the testing data.

We evaluate the final result using both the accuracy and F1 score. Accuracy measures all the correctly classified cases. However, accuracy alone is not sufficient for evaluation, because accuracy treats all different classes equally. All our datasets have notable class imbalances, so we also evaluate using the F-1 score, which is the harmonic mean of the precision and recall scores. The F1 score considers how the data is distributed and measures the incorrectly classified cases.

BERT, or Bidirectional Encoder Representations from Transformers, is proposed by Devlin et. al. in 2018 (Devlin et al., 2018). A Transformer is a neural network that maps every output element to every input element with regard to attention. This way it learns contexts by assigning attention to sequential data like sentences, thus being able to track relationships between each element like the words in a sentence. BERT is built on top of the Transformer model. It is designed to have bidirectionality, meaning that it will read text input in both left-to-right and right-to-left direction at the same time. This bidirectionality allows BERT to use the surrounding words to establish context.

HateBERT (Caselli et al., 2020) is a retrained BERT model with the specific task of abusive language detection. The model was trained on RAL-E, a Reddit comments dataset consisting of banned comments for being offensive, abusive, or hateful. It was trained with the BERT base-uncased model and the Masked Language Model (MLM) objective.

CyberBERT (Paul and Saha, 2022) is another BERT-based cyberbully detection model. The authors of CyberBERT added a fully connected layer over the final hidden state for cyberbully classification. They also further optimized the model with an additional softmax classifier during the fine-tuning phase.

3.3. Experimental Results

We ran the three models with the same three datasets, and we report the result in Table 2.

For the vanilla BERT model, we see that it performed much better on the Twitter dataset than YouTube and Kaggle. This is because the Twitter dataset has way significantly more data points than the other two, meaning that BERT received a lot more training data for fine-tuning when test-

[1] https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset/code

	F1	Accuracy
Twitter	0.705	0.692
YouTube	0.410	0.488
Kaggle	0.496	0.500

(a) BERT

	F1	Accuracy
Twitter	0.885	0.873
YouTube	0.816	0.803
Kaggle	0.797	0.771

(b) HateBERT

	F1	Accuracy
Twitter	0.849	0.861
YouTube	0.794	0.799
Kaggle	0.748	0.736

(c) CyberBERT

Table 2: Cyberbully detection model evaluations.

ing on the Twitter dataset. It also performed better on the Kaggle dataset than the YouTube dataset. We hypothesize that the reason behind this behavior is the imbalance in the dataset. Even though both datasets are imbalanced, the cyberbully to non-cyberbully data points in the YouTube dataset is remarkably higher than the ratio in the Kaggle dataset. The cyberbully to non-cyberbully data ratio of the YouTube dataset is 0.13, while the ratio for the Kaggle dataset is 0.46. Guo et al. explored this dataset imbalance in their paper published in 2022 (Guo et al., 2022). They proposed an architecture that first generates enough data so that the dataset is balanced, then fine-tune their model with the new augmented dataset. Their evaluation sees an improvement in the final result.

HateBERT and CyberBERT have similar performances, but HateBERT performed slightly better, so we choose to use HateBERT for our proposed model and future evaluation.

3.4. Failure Analysis

When we look at the misclassified cases, we observe that a lot of misclassified cases contain sarcasm. We provide examples of sarcasm in cyberbullying below:

- For the first time in my months of monitoring this, a man momentarily surpassed all the LWs in targeted GamerGate harassment. Congrats?
- 10% of the posts I've read on Facebook today are people looking for work. Jeez. I thought the unemployment rate was supposed to be better?
- i had a dream that i was once again being harassed by the girls who bullied me in high school. it was very vivid and accurate! i feel great about myself today

These sentences are taken from the Twitter dataset, and all three models classified them as non-cyberbullying. The experts who annotated the dataset considered it to be cyberbullying. These are false negative examples. On the opposite hand, we also observe non-cyberbullying sentences being classified into cyberbullying text, or false positive cases:

- I have learned that pleasing everyone is impossible, but pissing everyone off is easy and funny as f*ck!! #lovethatsh*t
- Hmmm. Perhaps some who are too pig-faced to get laid and therefore have zero chance of getting pregnant from such activity hold something against women who can?? IDK. Stream of consciousnees thought after looking at her.
- f*cking weird stupid game man, can't believe we still won

We hypothesize that the misclassifications are due to the use of irony, which according to the Oxford English Dictionary is defined as "the use of words to express something other than and especially the opposite of the literal meaning of a sentence". Sarcasm is a special case of irony that has a bitter, caustic tone that is "usually directed against an individual". We propose that irony affects both false positives and false negatives. In the false negative case, the aggressors may use words that appear innocent by definition, but the context suggests that the sentence is insulting due to sarcasm. Conversely in the false negative case, some words may appear hostile, but with context either the words are not used toward a specific person, or the hostile word is used as an irony. We hypothesize that integrating irony and sarcasm directly into our models will improve their cyberbullying classification performance.

It is worth noting that one of the main cues of sarcasm is the intonation of speech, thus detecting sarcasm by text alone can be challenging. Different people may have different judgment. However, one of the main components of sarcasm detection is context, and BERT is one of the best tools for understanding contextual cues within a sentence based on its bidirectionality. We believe that humans may disagree with the result produced by a sarcasm detection model, but the sarcasm model is sufficient for the purpose of cyberbully detection.

4. Proposed Method

4.1. Sarcasm Detection Layer

First we evaluate each sarcasm detection method. We use the dataset gathered by (Shmueli et al., 2020), which consists of 15,000 sarcastic and 15,000 non-sarcastic tweets.[2] We randomly chose 5,000 data points from each class for the testing dataset.

We use a neural network-based sarcasm detection model from (Ghosh and Veale, 2016). We do not re-train or fine-tune the model. We achieved an accuracy of 0.829 on our testing dataset, and we deem that sufficient for our purpose. The model uses a CNN-LSTM architecture, which converges faster than LSTM alone and produces a better composite representation of the input sentence. The dropout layer on top of the CNN was also removed, as the authors observed that some sarcasm indicator words were dropped out from the output of the CNN layer.

4.2. Multilayer Perceptron

The last layer of our model is a simple multilayer perceptron (MLP). The input consists of the HateBERT output, the BERT embedding of the input sentence, and the output from the sarcasm detection model. Note that BERT produces a larger embedding vector than HateBERT. When training the MLP, we trained two different models, one with BERT embedding and the other with HateBERT embedding. We find that both the training time and the accuracy are similar for the two models, and we conclude that the embedding method will not significantly affect the performance result. The output is the final cyberbully classifier. We have two hidden layers followed by an output layer. Accuracy metric is used in the training of the model, as we stop training the model when there is no more accuracy improvement for 15 epochs. We use sigmoid as our activation function and Adagrad as our optimizer.

We choose an MLP for our experiment because it is a weight-based network. During the training of MLP, it can identify the weight of each input feature. Using a more complex deep learning architecture may further improve the cyberbullying detection performance, which can be explored in future works.

4.3. Results

Similar to the cyberbully detection model evaluation, we use both the accuracy metrics and F1 score. The experimental results are reported in the table below:

	F1	Accuracy
Twitter	0.885	0.873
YouTube	0.816	0.803
Kaggle	0.797	0.771

(a) HateBERT

	F1	Accuracy
Twitter	0.937	0.924
YouTube	0.891	0.859
Kaggle	0.808	0.813

(b) HateBERT + Sarcasm

We see improvement in all three datasets. Note that the Twitter dataset has the most significant improvement. It is also the largest and the most imbalanced dataset among the three. The sarcasm detection model is also trained using a separate Twitter dataset, which may be one of the causes for the most improvement. However, we do see that the sarcasm detection improved the performance on the YouTube and Kaggle datasets. For our experiment purpose, we do not assume that the sarcasm detection model can correctly detect sarcasm, but rather output a feature score that plays a role in the final cyberbullying detection.

4.4. Ablation Study

We want to investigate if the sarcasm detection model helps improve the classification, or if the additional MLP is the cause for improvement, so we decided to train a similar MLP without including the sarcasm detection model score. The results are shown below:

	F1	Accuracy
Twitter	0.882	0.871
YouTube	0.818	0.805
Kaggle	0.800	0.769

We see no significant improvement in the ablation study model, which confirmed our hypothesis that the sarcasm detection model is the main source of improvement. However, we do see a slight increase in the evaluation metrics with the addition of the MLP, but including the MLP also increases the training time. It is also noted that

[2] Datasets and instructions can be found at https://github.com/bshmueli/SPIRS

training the MLP with or without the sarcasm detection model score does not increase the training time, and the runtime also stays consistent with the two versions of MLP.

5. Conclusion

In this work, we compare several benchmark methods for cyberbully detection. We then perform a failure analysis to investigate where the methods failed to classify the data points accurately, and we observe the common characteristic of misclassified cases to be sarcasm. We hypothesize that the cyberbully classifiers do not perform well on ironic texts, and by including a sarcasm score in the final classification, we can improve both the accuracy and F1 score. We do not assume that all cyberbullying texts are sarcastic, but we believe that many false negative and false positive cases contain sarcasm.

We conduct an evaluation of sarcasm detection models. We choose the best cyberbully detection model and the best sarcasm detection model to create a simple MLP that takes the cyberbully score, the sarcasm score, as well as the BERT representation of the original input data point and outputs a final cyberbully classification. We find that our model outperforms all benchmark cyberbully detection models.

Our finding suggests that cyberbully detection may involve other NLP tasks, including but not limited to sarcasm, sentiment and emotion analysis, or intent classification, etc. Future work can be done to evaluate how each task affects the performance of cyberbully detection.

5.1. Discussion

We note that there is a discrepancy between the definition of cyberbullying. Most literature we reviewed has a similar definition of cyberbullying, which we defined in the introduction. However, several works choose to distinguish between hate speech and cyberbullying. Those works define hate speech as general insulting to a group or a community, and cyberbullying as a form of personal attack. For example, an attack toward a specific social group is hate speech and not cyberbullying, and an attack toward a person belonging to a specific social group is cyberbullying but not hate speech. We choose to not investigate the difference between hate speech and cyberbullying, meaning that we treat those two similarly, but further work may be performed on the difference in the definition of hate speech and cyberbullying, which can potentially increase the accuracy from training the data by the specific definition group.

Similarly, there exist discrepancies when classifying sarcastic comments on social media. During the investigation, we often find ourselves disagreeing with the sarcasm classification results. The length of the input data and the lack of contextual information can also hinder sarcasm classification performance. Sarcasm detection is indeed a difficult task, and we do not claim that our model can achieve outstanding performance on this task. We simply use a sarcasm detection model to extract features from a different standpoint, and use that feature to aid us in cyberbullying detection.

5.2. Limitations

It is worth noting that all datasets used in this project are human-annotated, meaning that the classification may be biased based on each annotator's knowledge, cultural background, definition of terms, etc. Some datasets are also dated back to 2018, which may become obsolete due to how fast the internet has evolved. These datasets do not represent all forms of cyberbullying, meaning that the results do not necessarily reflect the generalizability of our method. Further testing is required to use our method outside the scope of public social media texts.

Furthermore, we did not test how accurate the sarcasm classifier is on the cyberbully dataset. Evaluating the accuracy of the sarcasm classifier in the cyberbully dataset requires the cyberbully dataset to be human-annotated, which is beyond the scope of this project. Future work is required to evaluate the sarcasm detection model against cyberbullying dataset. We do not reject the possibility that the sarcasm detector is not detecting sarcasm in the data, but rather detecting some underlying features with correlation to cyberbullying that is not detected by the cyberbully detection models.

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. OIA-1946391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. Bibliographical References

Mohammed Ali Al-Garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433–443.

Darkunde Mayur Ashok, Agrawal Nidhi Ghanshyam, Sayed Saniya Salim, Dungarpur Burhanuddin Mazahir, and Bhushan S Thakare. 2020. Sarcasm detection using genetic optimization on lstm with cnn. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–4. IEEE.

Vimala Balakrishnan, Shahzaib Khan, and Hamid R Arabnia. 2020. Improving cyberbullying detection using twitter users' psychological features and machine learning. *Computers & Security*, 90:101710.

Jennifer Bayzick, April Kontostathis, and Lynne Edwards. 2011. Detecting the presence of cyberbullying using computer software.

Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. 2014. Detecting online harassment in social networks.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Saravanan Chandrasekaran, Aditya Kumar Singh Pundir, T Bheema Lingaiah, et al. 2022. Deep learning approaches for cyberbullying detection and classification on social media. *Computational Intelligence and Neuroscience*, 2022.

Niladri Chatterjee, Tanya Aggarwal, and Rishabh Maheshwari. 2020. Sarcasm detection using deep learning-based techniques. *Deep Learning-Based Approaches for Sentiment Analysis*, pages 237–258.

Vikas S Chavan and SS Shylaja. 2015. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2354–2358. IEEE.

Maral Dadvar, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012a. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

Maral Dadvar, Roeland Ordelman, Franciska De Jong, and Dolf Trieschnigg. 2012b. Towards user modelling in the combat against cyberbullying. In *Natural Language Processing and Information Systems: 17th International Conference on Applications of Natural Language to Information Systems, NLDB 2012, Groningen, The Netherlands, June 26-28, 2012. Proceedings 17*, pages 277–283. Springer.

Maral Dadvar, Dolf Trieschnigg, and Franciska De Jong. 2014. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Advances in Artificial Intelligence: 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings 27*, pages 275–281. Springer.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska De Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35*, pages 693–696. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 11–17.

Suliman Mohamed Fati, Amgad Muneer, Ayed Alwadain, and Abdullateef O Balogun. 2023. Cyberbullying detection on twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction. *Mathematics*, 11(16):3567.

Ziyang Feng, Jintao Su, and Junkuo Cao. 2022. Bhf: Bert-based hierarchical attention fusion network for cyberbullying remarks detection. In *Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing*, pages 1–7.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 161–169.

Ashish Goel and Latika Gupta. 2020. Social media in the times of covid-19. *JCR: Journal of Clinical Rheumatology*, 26(6):220–223.

Xiaoyu Guo, Usman Anjum, and Jusin Zhan. 2022. Cyberbully detection using bert with augmented texts. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1246–1253. IEEE.

Jeffrey T Hancock. 2004. Verbal irony use in face-to-face and computer-mediated conversations. *Journal of Language and Social Psychology*, 23(4):447–463.

Sameer Hinduja and Justin W Patchin. 2008. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant behavior*, 29(2):129–156.

Akshi Kumar, Saurabh Raj Sangwan, Anshika Arora, Anand Nayyar, Mohamed Abdel-Basset, et al. 2019. Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. *IEEE access*, 7:23319–23328.

Amanda Lenhart, Kristen Purcell, Aaron Smith, and Kathryn Zickuhr. 2010. Social media & mobile internet use among teens and young adults. millennials. *Pew internet & American life project*.

Vicente J Llorent, Rosario Ortega-Ruiz, and Izabela Zych. 2016. Bullying and cyberbullying in minorities: Are they more vulnerable than the majority group? *Frontiers in psychology*, 7:1507.

Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. 2008. Detecting flames and insults in text.

Vinita Nahar, Xue Li, and Chaoyi Pang. 2013. An effective approach for cyberbullying detection. *Communications in information science and management engineering*, 3(5):238.

Sayanta Paul and Sriparna Saha. 2022. Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification. *Multimedia Systems*, 28(6):1897–1904.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Pradeep Kumar Roy and Fenish Umeshbhai Mali. 2022. Cyberbullying detection using deep transfer learning. *Complex & Intelligent Systems*, 8(6):5449–5467.

Stephen M Serra and Hein S Venter. 2011. Mobile cyber-bullying: A proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness. In *2011 Information Security for South Africa*, pages 1–5. IEEE.

Boaz Shmueli, Lun-Wei Ku, and Soumya Ray. 2020. Reactive Supervision: A New Method for Collecting Sarcasm Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2553–2559, Online. Association for Computational Linguistics.

Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

Sterling Stauffer, Melissa Allen Heath, Sarah Marie Coyne, and Scott Ferrin. 2012. High school teachers' perceptions of cyberbullying prevention and intervention strategies. *Psychology in the Schools*, 49(4):352–367.

Jason Wang, Kaiqun Fu, and Chang-Tien Lu. 2020. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1699–1708. IEEE.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, pages 1–6.