Investigating Text and Image Features for the Detection of Cyberbullying

JULIA HEINE, IU International University of Applied Sciences, Germany
KRISTINA SCHAAFF, IU International University of Applied Sciences, Germany
TIM SCHLIPPE, IU International University of Applied Sciences, Germany

A very serious issue throughout social media platforms and cultural groups on the internet is the phenomenon of cyberbullying, an emerging form of victimization brought about through the digital age [34]. In this paper, we investigate traditional and new features to detect cyberbullying in text messages which contain images. Our best *text-based cyberbullying detection system* achieves an accuracy of 66.7% on a balanced subset of the *MMSH150K* dataset [11]. The system uses a combination of *text vector features*, *semantic features*, *list lookup features*, *emoji features*, *error-based features*, and *AI feedback features*. Our best *image-based cyberbullying detection system*, which leverages *image text features*, obtains an accuracy of 52.5%. This shows that text vector features and image vector features, which are mostly used for this task in related work, do not necessarily lead to the best results and our new features contribute a part to improving the performance.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Information extraction**; Neural networks.

Additional Key Words and Phrases: Cyberbullying, Natural Language Processing, Classification, Text and Image Features

ACM Reference Format:
Julia Heine, Kristina Schaaff, and Tim Schlippe. 2024. Investigating Text and Image Features for the Detection of Cyberbullying. 1, 1 (July 2024), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The digital transformation has revolutionized our way of communication. In 2023, 4.88 billion people used social media [35] which is almost 60% of the global population. This has also raised a new way to intimidate or abuse other people in a virtual way, a method also referred to as cyberbullying [2]. According to Smith et al. [34], cyberbullying refers to an aggressive, intentional act or behavior by an individual or a group repeatedly using electronic forms of contact against a victim who finds it challenging to defend themselves [34].

To ensure fair treatment of users on internet platforms, it is important to quickly identify people who engage in cyberbullying and to quickly flag and delete texts that contain cyberbullying.

Since the investigation of different text and image features as well as their combination has not been done extensively, we focus on the evaluation of features that have not yet been analyzed in the related work but have potential for this use case.

Consequently, our contributions are as follows: We investigate three modalities to detect cyberbullying in text messages containing images: (1) Using text features, (2) using image features, and (3) using their combination.

Authors' addresses: Julia Heine, julia.heine@iubh.de, IU International University of Applied Sciences, Germany; Kristina Schaaff, kristina.schaaff@iu.org, IU International University of Applied Sciences, Germany; Tim Schlippe, tim.schlippe@iu.org, IU International University of Applied Sciences, Germany.

In the following section, we will describe related work regarding cyberbullying in text messages. In Section 3, we will explain the features that we investigated for the classification. The experiments and the results to detect cyberbullying in text messages will be presented in Section 4. We will conclude our work in Section 5 and suggest further steps.

2 RELATED WORK

In this section, we will describe the related work that tackles the detection of cyberbullying in text messages using text and image features plus their combination.

2.1 Cyberbullying Detection using Text Features

Cyberbullying detection in text is more investigated than in images or in their combination:

For example, Islam et al. [16] analyzed Decision Trees [42], Naive Bayes Webb [40], Support Vector Machines (SVMs) Hearst et al. [12], and Random Forests [6] with *text vector features* to classify Facebook and Twitter comments as *bullying* or *non-bullying*. The combination of SVMs and Term Frequency-Inverse Document Frequency (TF-IDF) [30] features outperformed Bag-of-Words [26] features and resulted in 79% precision on the Facebook comments and 80% on the Twitter comments.

Raj et al. [27] investigated 11 classification techniques including traditional machine learning approaches and neural networks, and seven types of feature extraction and embedding techniques. With 98.12% F1 score on the *Wikipedia Attack Dataset* [41] and 98.77% on the *Wikipedia Web Toxicity Dataset* [41], the combination of an SVM with TF-IDF unigram embeddings performed best.

Alkomah et al. [1] compared Naive Bayes Webb [40], Gradient Boosting Bentéjac et al. [4], XGBoost [33], Random Forest [6], K-Nearest-Neighbor Cover and Hart [8], Decision Trees [42], Convolutional Neural Network (CNN) [22], Long Short-Term Memory (LSTM) [13] and Bidirectional Encoder Representations from Transformers (BERT) [10] together with three features categories: (1) Mentioned user name, counts of capital letters, hashtags, and emojis, (2) TF-IDF word embeddings, (3) *Linguistic Inquiry and Word Count* (*LIWC*). Alkomah et al. [1] report that the highest performance was achieved with the combination of BERT and TF-IDF word embeddings.

Atoum [3] investigated the performances of SVMs and Naive Bayes in combination with different *text vector features*. The best result of 88.66% F1 score on their Twitter dataset was obtained using SVM and 4-gram text feature vectors.

Van Bruwaene et al. [38] analyzed SVMs, XGBoost and CNNs in combination with *text vector features* and *LIWC features*. Using XGBoost together with *LIWC* and *text vector features* performed best with 58.8% F1 score on their bullying dataset.

2.2 Cyberbullying Detection using Image Features

Compared to cyberbullying detection in texts, less focus has been on the detection in images. Some researchers considered image features separately, but almost all combined them with text features:

For example, Hosseinmardi et al. [14] investigated one-word descriptions derived from the images as features for a Naive Bayes and a maximum entropy (MaxEnt) classifier. Naive Bayes outperformed MaxEnt on their Instagram-based dataset of images and corresponding comments. With F1 scores of 49%, the authors claim that only focusing on visual data is not enough for cyberbullying detection.

Gomez et al. [11] used a Feature Concatenation Model (FCM) that consists of three fully-connected layers with decreasing dimensionality and ReLu activation functions to classify the images in the *MMHS150K* dataset, leading to 66.7% F1 score.

Botelho et al. [5] experimented with Xception [7], NASNet and Inception-ResNet V2 to classify the images of an MMHS150K subset into *bullying* and *non-bullying*. Xception performed best with 54.4% F1 score.

2.3 Cyberbullying Detection using a Text-Image Feature Combination

Vidgen et al. [39] explain that the combination of text and image information often reveals hateful content, which cannot be recognized by looking at the image or the text alone.

Hosseinmardi et al. [14] analyzed Naive Bayes and MaxEnt in the multi-modal model setting of text and image. MaxEnt in combination with image content, user properties, post time, and image caption as features performed best on their Instagram-based dataset with 68% F1 score.

Gomez et al. [11] combined image feature vectors derived from Google's Inception v3 [36], text vectors produced from GloVe [24] and image text vectors generated by the Text Detection module of Google's Vision API and GloVe in a Feature Concatenation Model (FCM), a Spatial Concatenation Model (SCM) and a Textual Kernels Model (TKM). The best result was achieved with the FCM, leading to 70.4% F1 score on the *MMHS150K* dataset.

Sahu et al. [29] analyzed two different fusion techniques, namely generative adversarial network (GAN)-Fusion and Auto-Fusion to model inter- and intra-modal dynamics and compared different classification models including FCM, SCM, TKM, uni-modal Bidirectional Long Short-Term Memory Network (BiLSTM) with attention and a multi-modal system that employed a Visual Geometry Group (VGG) to encode images and a BiLSTM with attention to encode text. Encompassing images, text, and captions through a concatenative fusion strategy produced highest F1 scores of 70% on the *MMHS150K* dataset.

Paul et al. [23] compared a recurrent neural network (RNN) based decision-level fusion to combine the models' outcomes and a stacking-based feature-level fusion where the feature information is combined before classification. The best option was to implement the feature-level fusion with a ResBiLSTM-RNN model leading to 75% F1 score.

3 OUR TEXT AND IMAGE FEATURES FOR CYBERBULLYING DETECTION

Since related work shows that *text vector features* usually outperform other feature categories, we also leverage *text vector features*. Furthermore, we focus on features that have not been analyzed in related work but have potential for this use case. In total, we implemented 18 text and image features which can be grouped into 9 feature categories. In addition to features which have already been studied in related work, we included 8 new features.

3.1 Text Features

We implemented the text feature categories *text vector features*, *semantic features*, *list lookup features*, *emoji features*, *error-based features*, and *AI feedback features*.

3.1.1 Text Vector Features. To classify the texts using their content, we analyzed *text vector features* which were successful in Hosseinmardi et al. [14], Gomez et al. [11], Islam et al. [16], Atoum [3], Van Bruwaene et al. [38],Botelho et al. [5], Sahu et al. [29], Raj et al. [27], Alkomah et al. [1], and Paul et al. [23]. To benefit from the semantic vector space, we used the Sentence-BERT implementation `bert-base-nli-mean-tokens`[1].

3.1.2 Semantic Features. *Semantic features* reflect the meaning in a text. We investigated the following *semantic features* to detect cyberbullying: sentiment polarity ($sentiment_{score}$), degree of subjectivity or objectivity ($subjectivity_{score}$),

[1] huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

degree of sarcasm ($sarcasm_{score}$), and degree of toxicity ($toxicity_{score}$). While $sentiment_{score}$ was successful in Desai et al. [9] and Perera and Fernando [25], we were the first to analyze $subjectivity_{score}$, $sarcasm_{score}$, and $toxicity_{score}$ for cyberbullying detection.

For the implementation of the $sentiment_{score}$ and $subjectivity_{score}$ features, we used the sentiment analysis of TextBlob[2]. For the implementation of the $sarcasm_{score}$ feature, we used the pre-trained BERT model from *textattack*[3]. For the generation of the $sarcasm_{score}$, we employed the pre-trained BERT model `toxic-bert`[4].

3.1.3 List Lookup Features. *List lookup features* provide information about the category of a word or character [21]. We investigated the following *list lookup features*: the relative frequency of capitalized words ($capitalizedChars_{rel}$), the relative frequency of special characters ($specialChars_{rel}$), and the relative frequency of special characters ($negWords_{rel}$). While $capitalizedChars_{rel}$ was successful in Huang et al. [15] and Alkomah et al. [1] and $negWords_{rel}$ was successful in Desai et al. [9], we were the first to analyze $specialChars_{rel}$ for cyberbullying detection.

3.1.4 Emoji Features. Emojis encapsulate a vast spectrum of emotions ranging from happiness and love to anger and sadness, often serving as markers of the underlying sentiment of a message [17]. Their capability to augment or replace words and convey sentiment makes them powerful tokens for analysis. Consequently and as they have been successful in Maity et al. [19], we investigated the following *emoji features*: $negEmoji_{score}$—indicating how negative the emojis are, $neuEmoji_{score}$—indicating how neutral the emojis are and $posEmoji_{score}$—indicating how positive the emojis are.

We derived the scores from the *Emoji Sentiment Ranking v1.0*[5] which contains 751 emoji characters with corresponding sentiment scores computed from 70,000 tweets [17].

3.1.5 Error-Based Features. The monitoring of linguistic errors could uncover behavioral patterns or emotional states of the writer. In the context of cyberbullying, these inaccuracies might be indicative of impulsive, aggressive, or emotionally charged messages. To the best of our knowledge, we are the first who analyze *error-based features* for cyberbullying detection.

For the detection of spelling and grammar errors, we used *LanguageTool*[6], also known as the spellchecker for OpenOffice.

3.1.6 AI Feedback Features. Another novel feature that—to the best of our knowledge—has not yet been used in the detection of cyberbullying is the *AI feedback feature*.

For this feature, we asked OpenAI's GPT-4 model through the *openai*[7] Python library if the text is hateful content. If GPT4 answered 'yes', we assigned the value 1 to the feature, otherwise 0.

3.2 Image Features

We investigated the image feature categories *image vector features*, *image text features*, and *image description features*.

3.2.1 Image Vector Features. To retrieve a feature representation of the whole image, we generated *image vectors features* using EfficientNet[8] [37].

[2] textblob.readthedocs.io/en/dev/quickstart.html
[3] huggingface.co/textattack/bert-base-uncased-SST-2
[4] huggingface.co/unitary/toxic-bert
[5] kt.ijs.si/data/Emoji_sentiment_ranking
[6] github.com/jxmorris12/language_tool_python
[7] pypi.org/project/openai/0.28.0
[8] keras.io/api/applications/efficientnet

3.2.2 Image Text Features. *Image text features* focus on the extraction and analysis of text embedded in images. As shown in Gomez et al. [11], cyberbullying detection models can greatly benefit from the inclusion of *image text features*, as they provide an additional layer of data that, when analyzed in conjunction with stand-alone text and image features, can provide a more nuanced understanding of the message.

To get a feature representation of the image texts provided in text form in the *MMSH150K* dataset, we converted the texts into sentence vectors using Sentence-BERT [28].

3.2.3 Image Description Features. *Image description features* represent the visible content of images by explaining the elements and scenarios depicted in images. We analyzed the following three image description features: *firstBestDescription*—one word representing the image's content, *firstBestDescriptionProbability*—the *firstBestDescription* together with the prediction probability and *fiveBestDescription*—the five words with the highest prediction probabilities.

To obtain 1-word image descriptions and corresponding probabilities, we first used *AlexNet* [18] from `torchvision`[9]. Then we converted the words into word vectors using GloVe [24].

4 EXPERIMENTS AND RESULTS

As proposed by [11], [29] and [23], we concatenated our feature vectors and trained a deep learning-based classifier. Since we achieved good results with multilayer perceptions (MLPs) in other text classification tasks [20, 31, 32], we chose to use an MLP to analyze stacked combinations of the features. To train, fine-tune, and test our systems and features, we used a subset of the *MMHS150K*[10] [11] dataset. We reduced the original number of six classes to the two classes of *bullying* and *non-bullying*. Since we found that we obtained worse results with the unbalanced training dataset, we took a subset of 29,469 messages with *bullying* and 29,469 messages with *non-bullying* from the original training set. We did not change the validation set (5,000 messages) or the test set (10,000 messages) defined by Gomez et al. [11]. Our analysis of this dataset's labels revealed that cyberbullying detection is a very hard task even for humans: All three annotators of the dataset agreed on whether it was *bullying* or *non-bullying* for only 46.38% of the labels. Only in 4.15% of the messages, all annotators agreed that it was *bullying*.

Tables 1, 2 and 3 show the accuracies (*Acc*) and F1 scores (*F1*) of our *text-based* and *image-based cyberbullying systems* which use an MLP in combination with the investigated text and image features plus their combinations. The best accuracy and the best *F1* are marked in bold. The systems which are based on our new features are marked with "*". Systems which combine only features which, when used alone, resulted in over 50% F1 score are labeled with "$_{F1>50\%}$".

Looking at the single text features in Table 1 shows that the *text vector features* (*Acc*=66.24%) perform best accuracy, followed by our new $toxicity_{score}$ features (*Acc*=59.61%) and the combination of our *semantic features* (*Acc*=59.04%). The combination of all text features ($text\ features_{all}$) slightly outperform the *text vector features* with 66.63% *Acc* but $Text\ Features_{F1>50\%}$ resulted in 66.68% *Acc*. In terms of *F1*, the best text features do not differ substantially: The best text features are $neuEmoji_{score}$ features (*F1*=66.39%), followed by our new $specialChars_{rel}$ features (*F1*=66.30%) and $negEmoji_{score}$ features (*F1*=66.29%). Neither $text\ features_{all}$ nor $text\ features_{F1>50\%}$ outperformed $neuEmoji_{score}$ in *F1*.

Analyzing the image features in Table 2 shows that only the *image text features* (*Acc*=52.54%), our new *firstBestDescriptionProb* features (*Acc*=50.16%), and *image vector features* (*Acc*=50.13%) lead to accuracies over 50%. In terms of *F1*, the *image vector features* (*F1*=65.27%) perform best, followed by $firstBestDescriptionProb$ (*F1*=62.37%) and the combination of *image description features* (*F1*=60.92%). The *F1* of the other image features is below 50%. Neither $image\ features_{all}$ nor $image\ features_{F1>50\%}$ outperformed the *image text features* in *Acc* or the *image vector features* in *F1*.

[9] pytorch.org/vision/0.8/models.html
[10] kaggle.com/datasets/victorcallejasf/multimodal-hate-speech

Table 1. Performances of Text Features plus their Combinations (* marks new features).

Feature Category	Acc	F1
Text Vector Features	66.24%	66.25%
Semantic Features	59.04%	56.24%
- $sentiment_{score}$	53.78%	39.92%
- $subjectivity_{score}$*	53.43%	43.43%
- $sarcasm_{score}$*	53.41%	53.21%
- $toxicity_{score}$*	59.61%	54.41%
List Lookup Features	55.52%	51.90%
- $capitalizedChars_{rel}$	51.09%	15.17%
- $negWords_{rel}$	53.94%	42.10%
- $specialChars_{rel}$*	50.29%	66.30%
Emoji Features	50.94%	56.53%
- $negEmoji_{score}$	50.67%	66.29%
- $neuEmoji_{score}$	50.42%	**66.39%**
- $posEmoji_{score}$	50.86%	65.54%
*Error-Based Features**	54.88%	56.76%
*AI Feedback Features**	56.06%	64.45%
Text $Features_{all}$	66.63%	65.54%
Text $Features_{F1>50\%}$	**66.68%**	66.08%

Table 2. Performances of Image Features plus their Combinations (* marks new features).

Feature Category	Acc	F1
Image Vector Features	50.13%	**65.27%**
Image Text Features	**52.54%**	31.57%
Image Description Features	49.96%	60.92%
- $firstBestDescription$	49.73%	25.11%
- $firstBestDescriptionProb$*	50.16%	62.37%
- $fiveBestDescription$*	49.89%	28.72%
Image $Features_{all}$	52.46%	30.39%
Image $Features_{F1>50\%}$	50.34%	56.10%

Table 3. Performances of the Combinations of Text and Image Features.

Feature Category	Acc	F1
Text $Features_{all}$	66.63%	65.54%
Text $Features_{F1>50\%}$	**66.68%**	66.08%
Image $Features_{all}$	52.46%	30.39%
Image $Features_{F1>50\%}$	50.34%	56.10%
Text+Image $Features_{all}$	64.19%	64.42%
Text+Image $Features_{F1>50\%}$	65.26%	65.37%

As demonstrated in Table 3, the combinations of text and image features did not outperform the best text features. This shows that our features can pull more information from text and the information we currently pull from images does not add value to improve *F1* and *Acc*.

5 CONCLUSION AND FUTURE WORK

We explored features for distinguishing between *bullying* and *non-bullying* text messages which contain images. Our best system with text features achieved 66.7% accuracy on the test set of the *MMSH150K* dataset. It employs a mix of *text vector*, *semantic*, *list lookup*, *emoji*, *error-based*, and *AI feedback* features. Our best system with image features obtained only 52.5% accuracy. Our investigation indicates that while the use of innovative text and image vector features does not lead to a substantial improvement in cyberbullying detection, given the inherently difficult nature of the task, these features can still provide valuable insights that may contribute to the development of future detection systems. Future work will include evaluating our features with other classifiers, on other datasets, and in other domains.

ACKNOWLEDGMENTS

This research was supported by the IU International University of Applied Sciences (*IU Incubator*) under the internal funding framework for the period from October 2023 to September 2025.

REFERENCES

[1] Fatimah Alkomah, Sanaz Salati, and Xiaogang Ma. 2022. A New Hate Speech Detection System based on Textual and Psychological Features. *International Journal of Advanced Computer Science and Applications* 13 (01 2022). https://doi.org/10.14569/IJACSA.2022.01308100

[2] Cambridge University Press Assessment. 2023. cyber. https://dictionary.cambridge.org/de/worterbuch/englisch/cyber

[3] Jalal Omer Atoum. 2020. Cyberbullying Detection Through Sentiment Analysis. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 292–297. https://doi.org/10.1109/CSCI51800.2020.00056

[4] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. 2021. A Comparative Analysis of Gradient Boosting Algorithms. *Artif. Intell. Rev.* 54, 3 (mar 2021), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[5] Austin Botelho, Scott Hale, and Bertie Vidgen. 2021. Deciphering Implicit Hate: Evaluating Automated Detection Algorithms for Multimodal Hate. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 1896–1907. https://doi.org/10.18653/v1/2021.findings-acl.166

[6] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[7] François Chollet. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. *CoRR* abs/1610.02357 (2016). arXiv:1610.02357 http://arxiv.org/abs/1610.02357

[8] T. Cover and P. Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27. https://doi.org/10.1109/TIT.1967.1053964

[9] Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, and Rashmi Dhumal. 2021. Cyber Bullying Detection on Social Media using Machine Learning. *ITM Web of Conferences* 40 (2021), 03038. https://doi.org/10.1051/itmconf/20214003038

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[11] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring Hate Speech Detection in Multimodal Publications. https://doi.org/10.48550/arXiv.1910.03814

[12] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support Vector Machines. *IEEE Intelligent Systems and their Applications* 13, 4 (1998), 18–28. https://doi.org/10.1109/5254.708428

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of Cyberbullying Incidents in a Media-Based Social Network. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Davis, California) *(ASONAM '16)*. IEEE Press, 186–192.

[15] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber Bullying Detection Using Social and Textual Analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia* (Orlando, Florida, USA) *(SAM '14)*. Association for Computing Machinery, New York, NY, USA, 3–6. https://doi.org/10.1145/2661126.2661133

[16] Md Manowarul Islam, Md Ashraf Uddin, Linta Islam, Arnisha Akter, Selina Sharmin, and Uzzal Kumar Acharjee. 2020. Cyberbullying Detection on Social Networks Using Machine Learning Approaches. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. 1–6. https://doi.org/10.1109/CSDE50874.2020.9411601

[17] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of Emojis. *PLOS ONE* 10 (12 2015), 1–22. https://doi.org/10.1371/journal.pone.0144296

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[19] Krishanu Maity, Sriparna Saha, and Pushpak Bhattacharyya. 2023. Emoji, Sentiment and Emotion Aided Cyberbullying Detection in Hinglish. *IEEE Transactions on Computational Social Systems* 10, 5 (2023), 2411–2420. https://doi.org/10.1109/TCSS.2022.3183046

[20] Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT. In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, Tim Schlippe, Eric C. K. Cheng, and Tianchong Wang (Eds.). Springer Nature Singapore, Singapore, 152–170. https://doi.org/10.1007/978-981-99-7947-9_12

[21] David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes* 30, 1 (Jan. 2007), 3–26. https://doi.org/10.1075/li.30.1.03nad

[22] Keiron O'Shea and Ryan Nash. 2015. An Introduction to Convolutional Neural Networks. arXiv:1511.08458

[23] Sayanta Paul, Sriparna Saha, and Mohammed Hasanuzzaman. 2022. Identification of Cyberbullying: A Deep Learning Based Multimodal Approach. *Multimedia Tools and Applications* 81 (2022), 1–20. https://doi.org/10.1007/s11042-020-09631-w

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[25] Andrea Perera and Pumudu Fernando. 2021. Accurate Cyberbullying Detection and Prevention on Social Media. *Procedia Computer Science* 181 (2021), 605–611. https://doi.org/10.1016/j.procs.2021.01.207 CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.

[26] Wisam A. Qader, Musa M. Ameen, and Bilal I. Ahmed. 2019. An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. In *2019 International Engineering Conference (IEC)*. 200–204. https://doi.org/10.1109/IEC47844.2019.8950616

[27] Chahat Raj, Ayush Agarwal, Gnana Bharathy, Bhuva Narayan, and Mukesh Prasad. 2021. Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques. *Electronics* 10, 22 (2021), 2810. https://doi.org/10.3390/electronics10222810

[28] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:201646309

[29] Gaurav Sahu, Robin Cohen, and Olga Vechtomova. 2021. Towards A Multi-Agent System for Online Hate Speech Detection. arXiv:2105.01129 [cs.AI]

[30] Gerard Salton and Chris Buckley. 1987. *Term Weighting Approaches in Automatic Text Retrieval*. Technical Report. USA.

[31] Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2023. Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. In *The 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)* (Virtual, 16-17 December 2023). https://aclanthology.org/2023.icnlsp-1.1.pdf

[32] Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2024. Classification of Human- and AI-Generated Texts for Different Languages and Domains. *International Journal of Speech Technology* (2024).

[33] Rexhep Shijaku and Ercan Canhasi. 2023. ChatGPT Generated Text Detection. (01 2023). https://doi.org/10.13140/RG.2.2.21317.52960

[34] 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, and allied disciplines* 49, 4 (2008), 376–385. https://doi.org/10.1111/j.1469-7610.2007.01846.x

[35] Statista Research Department. 2023. Number of Internet and Social Media Users Worldwide as of July 2023. https://www.statista.com/statistics/617136/digital-population-worldwide/

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[37] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networ. In *The 36th International Conference on Machine Learning* (Long Beach, California, USA).

[38] David Van Bruwaene, Qianjia Huang, and Diana Inkpen. 2020. A Multi-Platform Dataset for Detecting Cyberbullying in Social Media. *Lang. Resour. Eval.* 54, 4 (dec 2020), 851–874. https://doi.org/10.1007/s10579-020-09488-3

[39] Bertie Vidgen, Helen Margetts, and Alex Harris. 2019. How Much Online Abuse is There? A Systematic Review of Evidence for the UK. https://doi.org/10.5281/zenodo.3582599

[40] Geoffrey I. Webb. 2010. *Naïve Bayes*. Springer US, Boston, MA, 713–714. https://doi.org/10.1007/978-0-387-30164-8_576

[41] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1391–1399. https://doi.org/10.1145/3038912.3052591

[42] Wataru Zaitsu and Mingzhe Jin. 2023. Distinguishing ChatGPT(-3.5, -4)-Generated and Human-Written Papers Through Japanese Stylometric Analysis. arXiv:2304.05534 [cs.CL]