Classification of Text, Image and Audio Messages Used for Cyberbulling on Social Medias

Ermira Idrizi, Mentor Hamiti
South East European University, Faculty of Contemporary Sciences and Technologies, Tetovo, North Macedonia
e.idrizi@seeu.edu.mk, m.hamiti@seeu.edu.mk

***Abstract* - Cyberbullying has become an increasingly significant cultural issue in recent years. A person is affected by cyberbullying in both psychological and emotional ways. Bullying that takes place via electronic devices, such as a computer, a smartphone, or a tablet computer, is known as cyberbullying. Harassment in the digital world can take many forms, including but not limited to transmitting or uploading insulting, harmful, inaccurate, or offensive material about another person via text, message, or program; or on social networking sites, message boards, and online games. Disclosure of personal or sensitive information about another individual might be considered cyberstalking. This paper will analyze several media types (text, images, and videos) posted on social media with the goal of identifying instances of cyberbullying. In this study, a graph convolutional neural network, a pretrained Googlenet, a Mel-scale filter bank speech spectrogram, and a CNN network model are introduced for use in audio post-classification. The study's primary findings indicate that audio post-processing with MFCC's and graph convolutional neural networks generates improved outcomes, including one-dimensional representation, for both text and image properties. To this end, we used a combination of GCN and Melfrequency cepstrum to represent text, image, and video input in this setting, with a resulting 85% accuracy of a bullying class.**

Keywords – Cyberbullying, Image, Text, Audios, Facebook, Instagram

I. Introduction

As a result of the proliferation of social media platforms and their ever-increasing user bases, internet consumers now maintain an almost constant presence online. Users will post their thoughts, opinions, or feelings on social media, which will then be remarked on or addressed in public or private chats. In recent years, there has been a proliferation of brand new services that make it possible for users to view films, images, and comments pertaining to any kind of event. Many amusing pictures, videos, and comments have come out as a result of this trend. However, harmful videos, pictures, and comments have also been posted on social media. According to [1], some social networks, including Facebook, Twitter, and YouTube, are attempting to erase or prohibit articles that make it easier for people to become victims of cyberbullying. The ways in which individuals communicate with one another have undergone significant transformations over the past few decades, and this trend is expected to continue. Rapid technological progress has been brought about as a consequence of microaggressions that are not restricted to the external surroundings or facial expression settings. According to [2], the definition of conventional bullying is described as repeated and deliberate acts of violence directed toward a powerless victim. This power imbalance may be personal or emotional, and the assault may be verbal (for example, insults), physical (for example, hitting), emotional or contextual, or any combination of the three (e.g., gossiping). The term "harassment" can refer to anything from harmless teasing to overtly violent behavior, including but not limited to threats of self-harm, homicide, or other violent acts. From a different point of view, it's possible that overt and indirect forms of bullying are related. Direct bullying consists of assaults that are clearly visible on the victim, allowing the bully to be identified as the perpetrator. On the other hand, implicit abuse is transmitted to the person by a private entity (such as rumors), with the goal of isolating the victim so that the aggressor cannot be identified. This is done in order to prevent the victim from being able to identify the perpetrator of the abuse. With the advent of modern information and communication technologies (ICT) and the tremendous explosion of social network systems that allow teenagers to communicate online, social connections have gained a new mechanism for communication. This is particularly true in the case of online communication. Social interactions involving offensive online content are common in these online interactions, as this is one of the primary manifestations of violence [3] in cases of cyber harassment, such as cyberbullying. These online interactions take place in the context of social interactions. Abuse that occurs between peers through the use of electronic media can be described in a number of different ways; nevertheless, there is consensus among researchers that this type of behavior must be intentional, harmful, and repeated for it to be considered abuse.

A. *Cyberbullying forms*

Cyberbullying can take place via chat rooms, e-mail, instant chats, blogs, social networking sites, and texting. It can also take place using personal computers, laptops, mobile phones, and online video games. Verbal and emotional aggression are the most common forms of cyberbullying. This can take the form of things like spreading false rumors, disclosing personal information that is either genuine or erroneous via texts, or uploading things on social media websites. The use of fire, denigration, masquerading, outing and deceit, exclusion, and cheerful slapping are some of the other sorts of abuse. To prevent and respond to cyberbullying, digital solutions that include automatic cyberbullying identification and messages that encourage behavioral self-reflection can be utilized. Therefore, to reduce instances of cyberbullying, classifiers must be as precise as feasible. However, it is still difficult to detect these events online in near-real time, dependent on the operationalization of cyberbullying given by automated cyberbullying detection systems, among other things [4].

B. *Cyberbullying algorithm for detections*

Cyberbullying is a major issue in modern society. Unfortunately, perusing the entire Internet to uncover cyberbullying postings is like hunting for a needle in a haystack, while leaving cyberbullying victims exposed to negative words causes them to get depressed, self-mutilate, or commit suicide. Cyberbullying is addressed manually via Internet Patrol. Children and teens use Instagram and Facebook to post photos and videos. Visual (image and video) content accounts for over 70% of internet traffic, and over 80% of adolescents use mobile phones, making them the most ubiquitous technology and a common cyberbullying platform [8]. Deep learning can detect cyberbullying on social media, curation, wikis, tweeting, forums, and bookmarking. These algorithms automatically detect cyberbullying in massive data sets. These self-organizing and self-learning algorithms can help detect online abuse or bullying and automate cyberbullying detection. Computer scientists now use data collection methods to anticipate cyberbullying instead of language and graphics. Data mining is a new field that extracts useful information from massive amounts of data. Data scientists utilized deep learning algorithm to predict cyberbullying messages using picture, text, video, and URL data. Current cyberbullying detection methods either search social media for text messages or integrate text and visual attributes. The current study focuses on developing a deep learning model that predicts the effect of cyberbullying texts on the general sentiment of individuals [5]. Unique to the current technique is the consideration of the impact of sentiment analysis on cyberbullying identification. In addition, the effect of video and URL as attributes and how they can be used to forecast cyberbullying communications. A significant contribution of this study is the use of video and URL analysis in the building of a deep learning model to detect cyberbullying texts.

II. Related Studies

A. *Text-based cyberbullying detection*

The first work to detect bullying behaviors that include social media harassment [6], using the CAW 2.0 Twitter corpus, the Naive Bayes(NB) was trained, Sequential Minimal Optimization (SMO), J48, bagging, and digging algorithms using the information benefit algorithm. Utilize the Synthetic Minority Oversampling Technique to construct a sound training dataset (SMOTE). This model included textual and social network components. Twitter's messages are described as "1.5-ego networks." A network consists of nodes and ties. Links represent the number of postings shared between users, while nodes represent the users themselves [7]. This ego network is used to determine which nodes have no contact with their peers and which people are more influential inside their social networks. As a result of the increased number of messages sent between users, bullying messages were described. In the digging classification algorithm, the ROC curve and true positive rate yielded the best results. However, simply the quantity of messages exchanged was utilized to establish whether or not a post was harassing, and demographic factors were not considered in the study. A system was suggested that uses SNA (Social Network Analysis) and text mining to identify key cyberbullies from 650,000 posts collected through web crawling using three sets of features: relationship between users, positive/negative rate of comments, and insulting word rate of comments based on group level and user level with random forest, logistic regression, and SVM classifiers [8]. A random forest classifier gave them a precision of 0.81 on the consumer level and an adapted to specific ratio of 0.74:1 on the group level[9].

B. *Cyberbullying Detection: Applying Text and Pictures*

Today, image-based cyberbullying is widespread and has far-reaching repercussions on society. They appear to be rapidly spreading through social networks. A VGG-16 and CNN (Convolutional Neural Network) were proposed for picture and text feature extraction [10]. The accuracy, recall, and F1 scores were for Twitter 80%, Facebook 79%, and Instagram 78% data, respectively. A two-dimensional convolution comparison on images and one-dimensional convolution on texts was implemented, as well as TF-IDF vectorization to transform one-dimensional text into a two-dimensional square matrix and integrate image and text into a distinct two-dimensional matrix. Then, a single-layer CNN with a wide filter is compared to a multiple-layer CNN with fewer filters and a 68 percent weighted average F1-score. It was proposed using enough graphic and text weight to distinguish cyberbullying communications in future investigations [22][23][24].

Due to the interdependence of posts like messages, pictures, sounds, and videos, social media cyberbullying detection is challenging [11][12]. In Table 1 we list some cyberbullying detection methods:

TABLE I. CYBERBULLYING DETECTION METHODS

Obtained Method	Findings	Limitations
Using Synthetic Minority Oversampling, cyberbullying messages can be classified. Huang and colleagues (2014).	Social connections improve precision.	Demographics are avoided during detection.
SVM and KNN algorithms (Hani Nurrahmi et.al 2017)	Provides RBF Kernel and parameter settings for the most accurate recognition of cyberbullying text.	POS tagger may struggle with multiple languages.
Convolutional Neural Network with Shortcuts at the Character Level (Nijia Lu et.al 2019)	The focal loss function and smallest learning unit address class imbalance.	This only works on personality characteristics and is expensive for larger datasets.
Deep Neural Network Architecture(Saima sadiq et.al 2020)	Unigram, Bigram, and TFIDF accurately extract key features.	Single-feature analysis takes time and reduces accuracy.
CNN and pretrained fastText(Sandip Modha et.al 2019)	A browser-based platform detects and visualizes online hostility.	Different hostile and sensitive topic thresholds can affect accuracy.

III. Methodology

Data was collected from leading online social media platforms including Facebook and Instagram by employing search terms such as fury, impotence, ugly, and scary photographs. By searching Google for image, text, and audio queries, the needed datasets was able to be generated. A graph convolution neural network to classify cyberbullying was used.

A. Cyberbullying Investigation Techniques Using SNA

Social network analysis refers to the statistical and qualitative examination of a networking website (SNA)[13]. SNA monitors and organizes the flow of connections and changes in concepts and relationships with insight. Websites, computers, animals, people, groups, organizations, and nations are examples of simple and complex entities. Consequently, implementing SNA metrics could aid us in researching and understanding many aspects of this network.

Recognition of the *Society A*, where society is a subset of nodes in a graph that are strongly connected to one another but weakly correlated to nodes in the graph's other neighborhoods. A connection with others using sites such has Facebook, Instagram, and Twitter were established. In these social circles, relatives, classmates, coworkers, and others may be included. Sensing societies in a network is one of the most significant jobs of network analysis. There could be millions of nodes and edges in a huge network, such as an online social network. It becomes extremely difficult to identify communities in such networks. Therefore, strategies for detecting societies that can partition the network into distinct neighborhoods are necessary. In the context of the agglomerative approach, we begin with a graph consisting of nodes from the actual line but without edges. The edges are then added one by one to the graph, beginning with the "larger" edges and moving to the "smaller" ones. This edge strength, often referred to as weight, is calculable and represents a subset of the individuals who posted the bullying remark. Analyzeing cyberbullying communications in the network using [14] Girvan-Newman Algorithm for Community Detection.

In this instance, neighborhood identification is used to discover significant and influential themes on social media that involve a big number of persons, and then to investigate the existence of cyberbullying messages behind these significant topics. Figure 1 depicts the community size distribution across the network: The X axis represents the number of communities, while the Y axis represents the size of the communities, which follows a power-law distribution. Power law degree distribution functions determine the likelihood that a node **u** has the degree **du.** Power law distributions and other related variations have the characteristic of being substantially peeled.

Formally, a probability distribution with heavy tails approaches zero for absolute values more slowly than a probability function. This implies that events that are effectively "impossible" in a conventional exponential distribution are assigned a non-trivial probability mass in huge distributions. Due of the heavy tails of degree distributions, the probability of encountering a tiny number of assurance degree nodes is not zero. Figure 1 represents networks large number of nodes with low degrees and a small number of nodes with extremely high degrees.

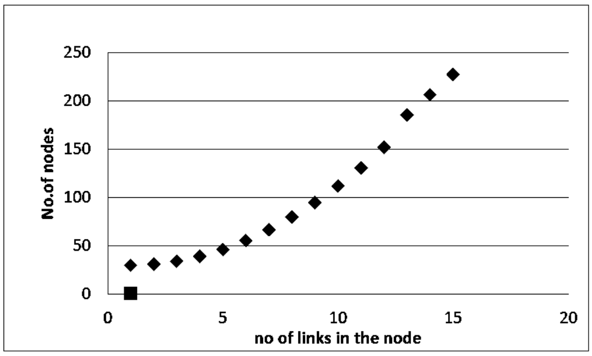

Figure. 1 Distribution of Power-law

B. Text Message Classification for Cyberbullying

It is commonly known that Twitter and Instagram are seeing an increase in cyberbullying towards young users, and that scalable computational solutions must be created to reduce this form of abuse.

Given this, a graph convolutional networks (GCN) was employed to identify cyberbullying [15]. The semi-supervised convolutional neural network-based GCN learns graph-structured data.

Node-level classification for graphs with few labels. This method accepts a feature matrix *ND* and feature description *xi* for each connection point *I*, where ***N*** is the number of vertices and ***D*** is the number of data quality. The amount of output features determines the node level output **Z N*F** from an adjacency matrix **AN*N**. Two dropout **GCN** layers with **L2** regularization were examined. Early Stopping with a 10-minute will be taught patience and 200 training epochs. If validation loss does not decrease for 10 epochs, training will stop. Transudative Learning means we'll train and test utilizing the full graph. Boolean masks divided training, validation, and testing data. These masks feed the sample weight argument. Before training a network, category features must be numeric. Single-pass vectorize categorical values.

Table 2 represents the results of GCC using Text data, with number of nodes 1240, average path length 2 with a F1 score 0.78 which leads to a accuracy of 85%.

TABLE 2. RESULTS OF GCC USING TEXT DATA

Nodes	Edges	Average Path Lenght	Average CC	Average Degree	F1 score	Accuracy
1240	28,975	2.02	0.221	1.876	0.78	85%

C. Classification of cyberbullying images

The initial component of the proposed model is GoogLeNet [16], a pretrained image model, a program for extracting photographic features. Using 3*3 filters, it applies the convolution layers to a 224*224*3 picture. Replace the last three layers of the network to retrain Google Photos to label recent photographs. The layers "loss3-classifier," "prob," and "output" include instructions for converting the network's characteristics into probabilities and symbols. To the layer graph, add three more layers: a fully connected layer, a softmax layer, and a classification output layer. The final, fully connected layer is the same size as the number of classes in the new data set, it increase the learning rate parameters of the entirely linked layer to learn new layers faster than transferred layers. Then "pool5-drop 7x7 s1" connects the newly transmitted network layer to the previously transmitted layer. A graph of the new layer is created and zoomed in on the final layers of the cable network to confirm that the new layers are properly connected. The network is now ready to be retrained with the new images. Optionally the weights are frozen of the network's older layers by setting their learning rates to zero.

During training, the network does not continuously update the variables of the frozen layers upon layers. Freezing the weights of several initial layers can considerably accelerate network training due to the elimination of gradient computations for the frozen layers. Freezing preceding network layers can protect against fitting problems if the new collection of data is minimal. By attempting to extract the layers and linkages from the layer graph, users can select which layers to freeze.

D. Classification of audio cyberbullying posts

Cyberbullying in audio form refers to bullying behavior that involves using audio recordings to harass, threaten, or intimidate someone. To classify audio cyberbullying posts, one could use several approaches:

- Content-based classification: This approach uses the content of the audio recordings to classify them as cyberbullying or not. Features such as speech content, tone, language, and sentiment could be used to train a machine learning model to automatically classify audio posts.

- Context-based classification: This approach considers the context in which the audio recording was made, such as the platform it was posted on, the identity of the speaker and the audience, and the relationship between the speaker and the target.

- Hybrid approach: A combination of content-based and context-based classification could also be used to get more accurate results.

Accurately identifying audio cyberbullying can be challenging, as the context of the audio and the intent of the speaker may not always be clear. Nevertheless, classification systems can be useful for detecting and addressing audio cyberbullying and for improving the overall safety and well-being of those affected by this type of behavior. The classification task for audio cyberbullying consists of two components: feature selection and categorization. A *Mel-scale filter* was applied to bank speech spectrogram as an input feature. The same signal was then shown in a time-frequency domain, allowing to evaluate the signal's time-varying frequency and amplitude. To achieve this result, a Fourier transform on our data was used. Cepstrum is the new name for the output spectrum, whereas Quefrencies refers to the new frequencies. With our new cepstrum and quefrequencies, we can easily discern between our spectral envelope (the vocal tract) and the background noise (the glottal pulse). Moreover, the discrete cosine transformation produces the mel-frequency cepstrum coefficients, or MFCCs, which are the major and most significant coefficients of this cepstrum. Figure 2 shows the main parts of a MFCC block diagram: frame blocking, windowing, FFT, Mel-frequeny and Cepstrum.

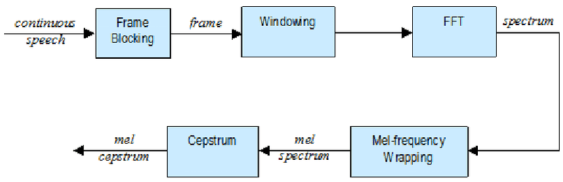

Figure. 2 MFCC Block Diagram[17]

In table 3 results are shown for text, images and voice using different models described apriority. Where all models show a high F1 score with accuracies all above 80%.

TABLE 3. RESULTS OF CLASSIFICATION USING DIFFERENT MODELS

Data	Model	F1 Score	Accuracy
Text	CNN	0.60	88%
	LSTM	0.61	89%
	1DCNN	0.63	91%
	GCNN	0.69	96%
Image	VGG-16	0.70	86%
	VGG-19	0.75	73%
	Resenet	0.78	75%
	GoogLeNet	.079	95.8%
Voice	MFCC	0.81	92%

E. Result analyzes

Using social network analysis, cyberbullying messages were categorized. To represent the data as a network, several SNA techniques were utilized. Using a graph-based deep learning method, the classification task was then performed. All existing methods accomplish the objective, however, using basic machine learning approaches and a limited data set. To assess and differentiate cyberbullying conversations, a novel whale optimization technique was applied on a wide variety of huge datasets, including text, image, and voice.

IV. Discussion and Conclusion

In addition to the benefits of social media, cyberbullying has a variety of well-established negative repercussions. Cyberbullying is one of the most damaging uses of social media, in which digital devices are used to threaten, intimidate, or humiliate an online user. Classification of cyberbullying messages using cutting-edge allometric data was established. Cyberbullying consequences and the degree of distribution and community size follow a power law distribution, and GoogLenet [18] and Mel-frequency cepstrum coefficients are employed to spread information. Each influential almetrics network community contains a number of cyberbullying messages that encourage the transmission of abusive or harsh comments about others as well as the sharing of videos, images, and texts with the intent to injure or dishonor others. Graph convolutional neural networks, GoogleNet, and MFCCs to develop an autonomous classification system for cyberbullying were employed. Our proposed method is capable of accurately categorizing 83% of cyberbullying scenarios. We discovered that GCN with Mel-frequency cepstrum coefficients and a one-dimensional convolution layer with a larger filter size beats multiple convolution layers with fewer filters [19].

For the cyberbullying identification test, only image, text, and voice were considered; however, video and the URL of the post can also be utilized to identify cases of cyberbullying. For the goal of cyberbullying identification, other characteristics of the post, such as network information, URL, and video content, may also be evaluated [20].

Images with multilingual annotations can be used to generate improved futures. To improve the overall effectiveness of cyberbullying comment detection, more optimization strategies for enhanced feature extraction must be investigated.

In conclusion, classifying cyberbullying messages on social media is a crucial task for ensuring the safety and well-being of individuals in online communities. By using NLP and machine learning techniques, it is possible to accurately identify and categorize messages that contain abusive or threatening language. This helps to mitigate the harm caused by cyberbullying and promote a safer and more positive online experience for users. However, it's important to note that this is an ongoing challenge and requires continuous improvement and refinement of these algorithms as cyberbullies adapt and find new ways to harass and harm others online.

References

[1] A. Sánchez-Medina, Galván-Sánchez I, Fernández-Monroy M. Applying artificial intelligence to explore sexual cyberbullying behaviour. Heliyon. 2020 Jan 1;6(1):e03218.

[2] R. Slonje, Smith PK, Frisén A. The nature of cyberbullying, and strategies for prevention. Computers in human behavior. 2013 Jan 1;29(1):26-32.

[3] GW. Giumetti, Kowalski RM. Cyberbullying via social media and well-being. Current Opinion in Psychology. 2022 Feb 19:101314.

[4] A. Bozyiğit , Utku S, Nasibov E. Cyberbullying detection: Utilizing social media features. Expert Systems with Applications. 2021

[5] M. Alotaibi, Alotaibi B, Razaque A. A multichannel deep learning framework for cyberbullying detection on social media. Electronics. 2021 Oct 31;10(21):2664.

[6] A. Perera, Fernando P. Accurate cyberbullying detection and prevention on social media. Procedia Computer Science. 2021 Jan 1;

[7] S.Bharti, Yadav AK, Kumar M, Yadav D. Cyberbullying detection from tweets using deep learning. Kybernetes. 2021 Jul 13;51(9).

[8] J. Zhang, Otomo T, Li L, Nakajima S. Cyberbullying detection on twitter using multiple textual features. In2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST) 2019 Oct 23 (pp. 1-6). IEEE.

[9] V. Balakrishnan, Khan S, Fernandez T, Arabnia HR. Cyberbullying detection on twitter using Big Five and Dark Triad features. Personality and individual differences. 2019.

[10] HC. Chan, Wong DS. Traditional school bullying and cyberbullying in Chinese societies: Prevalence and a review of the

whole-school intervention approach. Aggression and Violent Behavior 2015.

[11] J. Yadav, Kumar D, Chauhan D. Cyberbullying detection using pre-trained bert model. In2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020 Jul 2 (pp. 1096-1100). IEEE.

[12] M. Makram, Ali N, Mohammed A. Machine Learning Approach for Diagnosis of Heart Diseases. In2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) 2022 May 8 (pp. 69-74). IEEE.

[13] L. Cheng, Guo R, Silva Y, Hall D, Liu H. Hierarchical attention networks for cyberbullying detection on the instagram social network. InProceedings of the 2019 SIAM international conference on data mining 2019 May 6 (pp. 235-243). Society for Industrial and Applied Mathematics.

[14] L. Cheng, Mosallanezhad A, Silva YN, Hall DL, Liu H. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. InThe Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP) 2021 Aug (Vol. 1).

[15] F. Elsafoury, Katsigiannis S, Pervez Z, Ramzan N. When the timeline meets the pipeline: A survey on automated cyberbullying detection. IEEE access. 2021 Jul 21.

[16] D. Soni, Singh VK. See no evil, hear no evil: Audio-visual-textual cyberbullying detection. Proceedings of the ACM on Human-Computer Interaction, 2018.

[17] JS. Raj, Anantha Babu S. Smart Cyberbullying detection with Machine Learning. InDisruptive Technologies for Big Data and Cloud Applications: Proceedings of ICBDCC 2021 2022 Aug 2 (pp. 237-248).

[18] GB. Anwar, Anwar MW. Textual Cyberbullying detection using Ensemble of Machine Learning models. In2022 International Conference on IT and Industrial Technologies (ICIT) 2022 Oct 3 (pp. 1-7). IEEE.

[19] J. Qiu, Moh M, Moh TS. Multi-modal detection of cyberbullying on Twitter. InProceedings of the 2022 ACM Southeast Conference 2022 Apr 18 (pp. 9-16)

[20] PK. Roy, Mali FU. Cyberbullying detection using deep transfer learning. Complex & Intelligent Systems. 2022 Dec;8(6):5449-67.

[21] S. Bharadwaj, R., Kuzhalvaimozhi, S., & Vedavathi, N. (2023). A Novel Multimodal Hybrid Classifier Based Cyberbullying Detection for Social Media Platform. In Data Science and Algorithms in Systems: Proceedings of 6th Computational Methods in Systems and Software 2022, Vol. 2 (pp. 689-699). Cham: Springer International Publishing.

[22] W.M Yafooz., Al-Dhaqm, A., & Alsaeedi, A. (2023). Detecting Kids Cyberbullying Using Transfer Learning Approach: Transformer Fine-Tuning Models. In Kids Cybersecurity Using Computational Intelligence Techniques (pp. 255-267). Cham: Springer International Publishing.

[23] Y. Zhao, Chu, X., & Rong, K. (2023). Cyberbullying experience and bystander behavior in cyberbullying incidents: The serial mediating roles of perceived incident severity and empathy. Computers in Human Behavior, 138, 107484.