Smart Cyber Victimization Discovery on Twitter

Niloufar Shoeibi[1](✉), Nastaran Shoeibi[2], Vicente Julian[3], Sascha Ossowski[4], Angelica González Arrieta[5], and Pablo Chamoso[1]

[1] BISITE Research Group, University of Salamanca, Salamanca, Spain
Niloufar.Shoeibi@usal.es
[2] Babol Noshirvani University of Technology, Babol, Iran
[3] Department of Computer Systems and Computation, Universitat Politècnica de València, València, Spain
[4] Universidad Rey Juan Carlos, Madrid, Spain
[5] Department of Computer and Automation, University of Salamanca, Salamanca, Spain
https://bisite.usal.es/

Abstract. The advancement of technologies, the promotion of smartphones, and social networking have led to a high tendency among users to spend more time online interacting with each other via the available technologies. This is because they help overcome physical limitations and save time and energy by doing everything online. The rapid growth in this tendency has created the need for extra protection, by creating new rules and policies. However, sometimes users interrupt these rules and policies through unethical behavior. For example, bullying on social media platforms is a type of cyber victimization that can cause serious harm to individuals, leading to suicide. A firm step towards protecting the cyber society from victimization is to detect the topics that trigger the feeling of being a victim. In this paper, the focus is on Twitter, but it can be expanded to other platforms. The proposed method discovers cyber victimization by detecting the type of behavior leading to them being a victim. It consists of a text classification model, that is trained with a collected dataset of the official news since 2000, about suicide, self-harm, and cyberbullying. Results show that LinearSVC performs slightly better with an accuracy of 96%.

Keywords: Twitter · Cyberbullying · Suicide and self-harm · Cyber victim · Text classification · Text feature extraction

1 Introduction

Technological advancements, the popularity of online social networking sites, and having internet access, all contribute greatly to the quality of life but also have some ill effects, such as cyberattacks, cybercrimes, and cyberbullying. Therefore,

J. M. Corchado and S. Trabelsi (Eds.): SSCTIC 2021, LNNS 253, pp. 289–299, 2022.
https://doi.org/10.1007/978-3-030-78901-5_25

cybersecurity is a crucial matter for researchers, and detecting cyberbullies will lead to improving people's mental health and to making social networking sites safer [1].

Bullying is a repetitive, aggressive behavior that includes physical, verbal and social intimidation. Cyberbullying appeared as a new way of bullying and aggression with the use of digital technologies and can take place on social media and messaging platforms [2]. It appears as harassment, cyberstalking, cyberthreats, happy slapping, impersonation and denigration etc., and can lead to various health issues including mental, emotional and physical problems alongside face-to-face bullying. It has serious effects on both the victim and the aggressor. People who are bullied tend to be more insecure, can't concentrate, have depression, anxiety, self-harm, and even suicidal thinking and attempts. People who bully are more likely to abuse and harm others, do drugs and have behavioral issues [3].

Anyone can be a victim of cyberbullying, so it is important to identify it and report it to stop the cyber victimization. Social network platforms are trying their best to detect cyberbullying by improving their features and privacy policies. Even though there are lots of difficulties in implementing cyberbullying detection tools because of the human behavior is stochastic, and arbitrary, there are a lot of factors affecting the behavior of a person, the lack of datasets [4].

The proposed method focuses on discovering cyberbullying and prevent the future consequences and serious issues, such as self-harm and suicide attempts, in order to guarantee a peaceful and safe cybersociety. There are challenges in detecting cyberbullying; manual detection is time consuming, requires human involvement and is frustrating. There are few datasets available for this purpose, so most of them are labeled manually and because of the limited length of the tweets on twitter, only 168 characters can be used therefore it makes detection more difficult. Due to the lack of datasets and the need for a more complete dataset, a dataset has been created on the basis of the official news using official Google News API [5]since 2000 (New York Times news, etc.). The proposed architecture consists of two modules. One downloads the tweets from the Twitter platform and the other one is a text classification model which detects if the input text (tweet) is related to cyberbully, self-harm, and suicide with the accuracy of 96%.

This paper has been organized as follows: In Sect. 2, the related work is reviewed. Then, in Sect. 3, the overview and the architecture of the proposed method are presented. Finally, the results, conclusion and future work are discussed in Sect. 4.

2 Review of the State of the Art

In the field of user behavior mining on social media platforms, many studies have been carried out [6–9] and still, many doors are open to researchers in this area, to discover greater knowledge about human behavior in many different situations. In this paper, the focus is on cyber victimization detection and prevention, especially on Twitter, which is the most news-friendly social media platform and is the main target for investigating cyberbullying and the related psychological issues.

V. Balakrishnan et al. in [10] proposed a detection method to reduce cyberbullying in the basis of Twitter users' psychological characteristics like feelings and personalities. Users' personalities defined with Big Five and Dark Triad models, then they used machine learning classifiers like Naïve Bayes, Random Forest, and J48 for classifying tweets into four sections: bully, aggressor, spammer, and normal tweets. Results show that analyzing traits like extraversion, agreeableness, neuroticism, and psychopathy has a great impact on identifying online bullies.

In [11], researchers proposed techniques for the detection of cyberbullying and presented a comparative analysis, classifying multiple methods for cyberbullying detection. Many of them use the SVM classifier and have illimitable results, and one method used the unsupervised approach, and it is more complicated. The identification of cyber aggression is an essential factor in predicting cyberbullying, and user profile legitimacy detection plays a significant role in it.

In future smart cities [12–15] as well as the current physical world, issues such as bullying, harassment, and hate speech must be counteracted. Kumari et al. used the contents of social media to identify the cyberbullies in texts and images. It explains the single-layer convolutional Neural Network has better results with a unified representation of both text and image. Using text as an image is a more suitable model for data encoding. They applied three layers of text and three layers of a color image to interpret the input that presents a recall of 74% of the bullying class with one layer of Convolutional Neural Network [16].

Many studies concentrate on improving the cyberbullying detection performance of machine learning algorithms, as proposed models cause and strengthen unintended social biases. O. Gencoglu et al. in [17] introduced a model training method that uses fairness constraints and operates with different datasets. The result shows that varieties of unintended biases can be successfully mitigated without reducing the model's quality.

Muneer et al. in [18] applied seven different machine learning classifiers, namely, Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB), and Support Vector Machine (SVM) on a global dataset of 37,373 tweets from Twitter to detect cyberbullying without affecting the victims. These algorithms use accuracy, precision, recall, and F1 score as performance factors to conclude classifiers' recognition rate applied to the global dataset. Results indicate LR has a median accuracy of around 90.57%. Logistic regression obtained the best F1 score (0.928), SGD obtained the best precision (0.968), and SVM has the best recall (1.00).

Most of the existing cyberbullying detection techniques are supervised by a human and take time. However, Cheng et al. in [19] introduced an unsupervised cyberbullying detection model that has better performance than supervised models. This model includes two components: (1) a representation learning network for encoding social media using multi-modal features and (2) a multi-task learning network that identifies the bullies with a Gaussian Mixture Model. Their proposed model optimizes the parameters of both components for getting the liabilities of decoupled training.

Z. Abbass et al. in [20] proposed a three module framework: data preprocessing, classifying model builder, and prediction. For data classification, Multinomial Naïve Bayes (MNB), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) are used to build the prediction model. These algorithms achieve the precision, recall, and F-measure above 0.9. Also, the support vector machine performed better. This system has better accuracy than the existing network-based feature selection approach.

Balakrishnan et al. in [21]showed the relationship between personality traits and cyberbullying and introduced a way to detect cyberbullying by defining Big five and Dark Triad features. For cyberbullying classification, they used the Random Forest algorithm combined with a baseline algorithm including some Twitter features (i.e. amount of mentions, amount of followers and following, reputation, favorite count, status count, and the number of hashtags). Big Five and Dark Triad are notable in finding bullies, obtaining up to 96% (precision) and 95% (recall).

Automatic cyberbullying detection may help stop harassment and bullies on social media, using manually engineered features. Sadiq et al. in [22] applied multilayer perceptron and analyzed the state-of-the-art combination of CNNLSTM and CNN-BiLSTM in the deep neural network. This model identifies cyber harassments with 92% accuracy.

In Table 1, the summary of the selected papers related to social media user behavior mining focusing on cyberbully detection is presented, including the method proposed in each paper.

Table 1. The review of the state of the art on social media user behavior mining by focusing on the sub area of cyberbully detection.

Paper title	Area	Sub-area	Methodology
Improving cyberbullying detection using Twitter users' psychological features and machine learning[10]	Social Media Behavior Mining	Cyberbully Classification	Naïve Bayes, Random Forest, and J48
Taxonomy of Cyberbullying Detection and Prediction Techniques in Online Social Networks [11]	Social Media Behavior Mining	Cyberbully Detection	SVM
Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach [16]	Social Media Behavior Mining	Cyberbully Detection	CNN
Cyberbullying Detection with Fairness Constraints [17]	Social Media Behavior Mining	Cyberbully Detection	Fairness constraints
A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter [18]	Social Media Behavior Mining	Cyberbully Detection	LR, LGBM, SGD, RF, ADB, NB, and SVM
Unsupervised cyberbullying detection via time-informed gaussian mixture model	Social Media Behavior Mining [19]	Cyberbully Detection	Gaussian Mixture Model
A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning [20]	Social Media Behavior Mining	Cyberbully Detection	MNB,KNN, and SVM
Cyberbullying detection on twitter using Big Five and Dark Triad features [21]	Social Media Behavior Mining	Cyberbully Detection	Random Forest
Aggression detection through deep neural model on Twitter [22]	Social Media Behavior Mining	Cyberbully Detection	Combination of CNN-LSTM and CNN-BiLSTM

3 The Proposed Architecture for Cyber Victimization Detection

As has been discussed in the previous sections, finding cAs has been discussed in the previous sections, finding cyberbullying victims is crucial to stop self-harm and suicide attempts. It can help public organizations guarantee a safe cyber society by discovering the victims. Building a trustable dataset to solve this problem, has a significant value. For this reason, the official news released since 2000 has been taken into account. Figure 1 represents the distribution of the data by different Media channels.

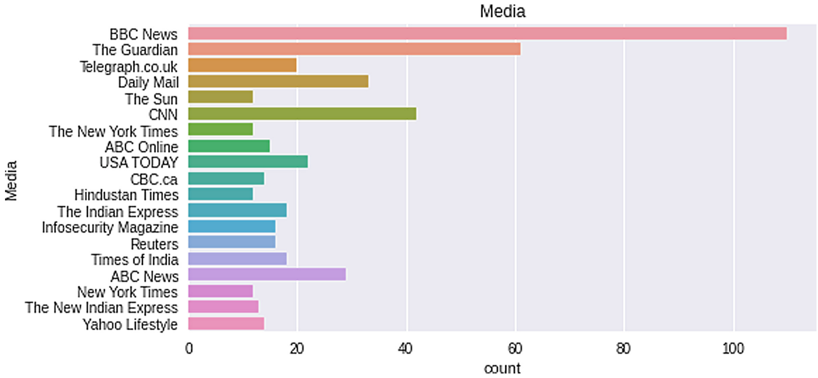


Fig. 1. Distribution of the data provided by official News Media.

Figure 2, represents the distribution of the data in the two different classes, also, it is understandable that the dataset is balanced, within the total of 1624 unique news articles.

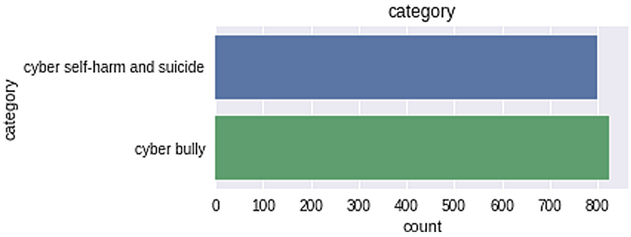


Fig. 2. Distribution of the data in the two categories.

The architecture proposed for cyber victimization detection has been presented in Fig. 3. This model consists of different stages, as discussed below.

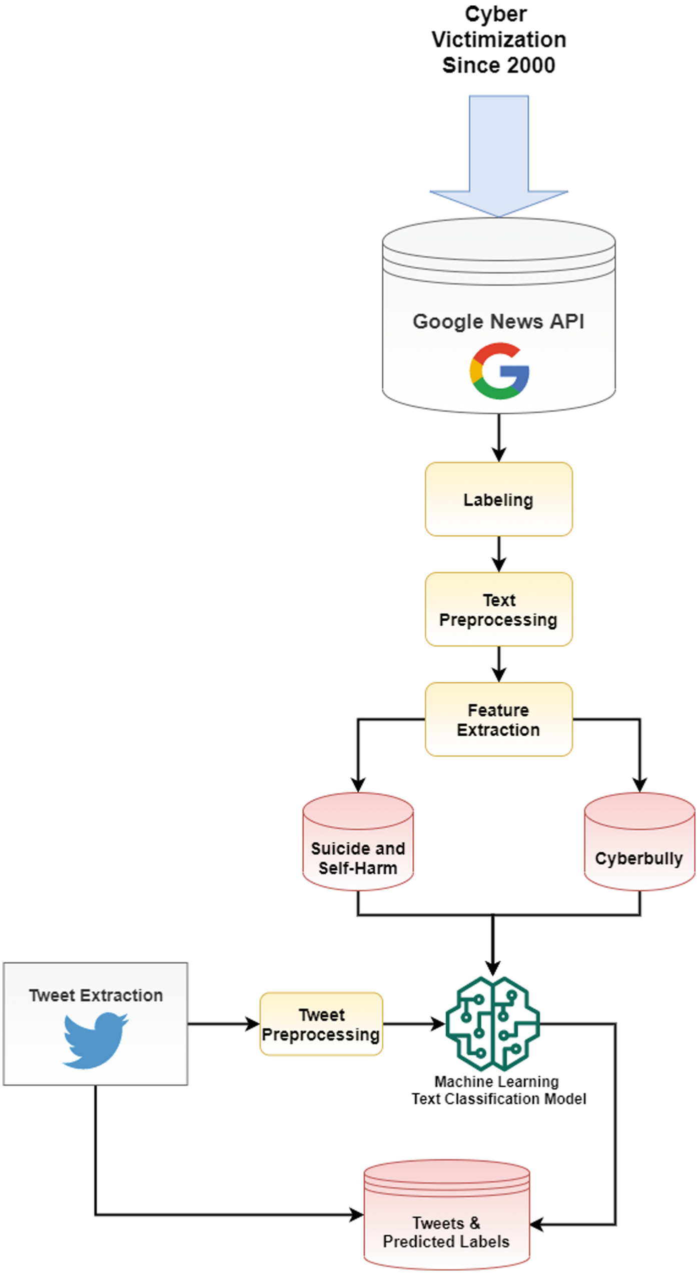


Fig. 3. The Architecture of Cyber Victimization Discovery.

First, a list of queries related to cyber victimization has been created, and using Google News official API, the news items related to each query are identified and stored for further processing. Then, these news items go through the procedure of labeling based on the topic of each article. The labeling component divides the data into two classes of news related to "CyberBully" and "Self-harm and Suicide" which is more general and also more urgent. If a user is posting about harming themselves, they have more priority.

After the labeling, the text of the articles goes through preprocessing including, tokenization which detects the words in the sentences, removing stopwords in English and 10 most frequent words, lemmatization which is the act of extracting the simple root of a word and then merging the tokens (preprocessed words) to create the cleaned text of each news article. Next, the clean text goes through the text feature extraction methods like count vectorizer and Tf-idf. Then, the dataset is shuffled and divided into train and test datasets.

In the end, three machine learning models are trained with this dataset and the model with the highest accuracy is selected to be used for the further steps.

As the aim of the model is to detect cyber victimization on Twitter, a query is done within the scope of the problem and the tweets are saved in a database ready to be processed. First, each tweet is preprocessed including tokenization, translation if the language of the tweet is not English, spell check, removing stopwords, and lemmatization. A spell check is necessary because on Twitter, due to the character limitation (168 characters), users tend to compact the words to be able to include more information in the tweet. After all these steps, the cleaned text is given to the Linear SVC model so that the label can be predicted. The labels are "cyberbully" or "cyber self-harm and suicide."

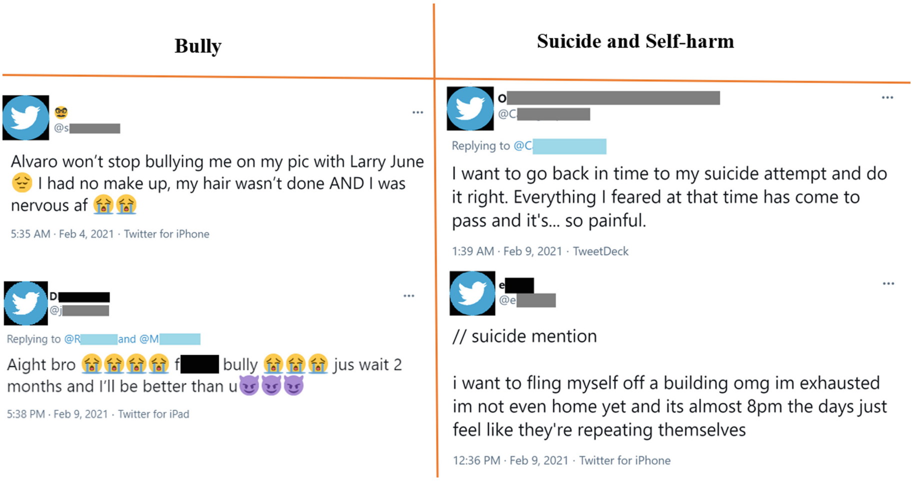
4 Results, Conclusion and Future Work

As has been explained in the previous section, three different classification models have been tried, the results have been presented in Table 2. It shows that the Linear SVC model has slightly better accuracy in comparison to the other two models.

Table 2. The classification report of Linear SVC, SGD, and Random Forest Classifier.

Model name	Classes	Precision	Recall	F1-score	Accuracy
LinearSVC	Cyberbully	0.96	0.95	0.96	0.96
	Self-harm & suicide	0.96	0.97	0.96	
SGD	Cyberbully	0.97	0.94	0.95	0.9569
	Self-harm & suicide	0.95	0.97	0.96	
Random Forest Classifier	Cyberbully	0.97	0.93	0.95	0.9507
	Self-harm & suicide	0.94	0.97	0.95	

Here are some samples of the tweets, completely anonymized that have been detected with the bully and self-harm and suicide related content presented in Fig. 4.



Bully

🤓
@s[redacted]

Alvaro won't stop bullying me on my pic with Larry June 😔 I had no make up, my hair wasn't done AND I was nervous af 😭😭

5:35 AM · Feb 4, 2021 · Twitter for iPhone

D[redacted]
@j[redacted]

Replying to @R[redacted] and @M[redacted]

Aight bro 😭😭😭😭 f[redacted] bully 😭😭😭 jus wait 2 months and I'll be better than u😈😈😈

5:38 PM · Feb 9, 2021 · Twitter for iPad

Suicide and Self-harm

o[redacted]
@C[redacted]

Replying to @C[redacted]

I want to go back in time to my suicide attempt and do it right. Everything I feared at that time has come to pass and it's... so painful.

1:39 AM · Feb 9, 2021 · TweetDeck

e[redacted]
@e[redacted]

// suicide mention

i want to fling myself off a building omg im exhausted im not even home yet and its almost 8pm the days just feel like they're repeating themselves

12:36 PM · Feb 9, 2021 · Twitter for iPhone

Fig. 4. The labeled tweet samples.

In this paper, a model has been presented that helps public organizations discover cyber victimization. The motivation behind building this model is to help the victims of cyberbullying and the victims who mention harming themselves or committing suicide. The architecture includes a text classification model whose accuracy is 96% (Linear SVC model) that has been trained with the official news published since 2000 within the scope of the cyber victimization by using count vectorizer and Tf-idf. Then, a query on Twitter is done by utilizing the official Twitter APIs. Later, tweets are preprocessed and after cleaning the tweets, this model is used to predict the label of the tweets related to this subject.

In the future, the plan is to expand the boundaries by identifying the aggressors as well as the victims, by reviewing the timeline and their activities to help them and motivate them to maintain a healthier lifestyle. Moreover, in the future the proposal will be implemented on the deepint.net platform which supports all types of data and contains a full suite of artificial intelligence techniques for data analysis, including data classification, clustering, prediction, optimization, and visualization techniques [23]. These abilities make it a perfect choice for implementing the proposal.

Acknowledgments. This research has been supported by the project "Intelligent and sustainable mobility supported by multi-agent systems and edge computing (InEDGE-Mobility): Towards Sustainable Intelligent Mobility: Blockchain-based framework for IoT Security", Reference: RTI2018-095390-B-C31/32/33, financed by the Spanish Ministry of Science, Innovation and Universities (MCIU), the State Research Agency (AEI) and the European Regional Development Fund (FEDER).

References

1. Bussey, K., Luo, A., Fitzpatrick, S., Allison, K.: Defending victims of cyberbullying: the role of self-efficacy and moral disengagement. J. School Psychol. **78**, 1–12 (2020)
2. Smith, P.K.: Research on cyberbullying: strengths and limitations. In: Vandebosch, H., Green, L. (eds.) Narratives in Research and Interventions on Cyberbullying Among Young People, pp. 9–27. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-04960-7_2
3. Martínez-Monteagudo, M.C., Delgado, B., Díaz-Herrero, Á., García-Fernández, J.M.: Relationship between suicidal thinking, anxiety, depression and stress in university students who are victims of cyberbullying. Psychiatry Res. **286**, 112856 (2020)
4. Raza, M.O., Memon, M., Bhatti, S., Bux, R.: Detecting cyberbullying in social commentary using supervised machine learning. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) FICC 2020. AISC, vol. 1130, pp. 621–630, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39442-4_45
5. Google news API
6. Shoeibi, N., Mateos, A.M., Camacho, A.R., Corchado, J.M.: A feature based approach on behavior analysis of the users on Twitter: a case study of AusOpen tennis championship. In: Dong, Y., Herrera-Viedma, E., Matsui, K., Omatsu, S., González Briones, A., Rodríguez González, S. (eds.) DCAI 2020. AISC, vol. 1237, pp. 284–294, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-53036-5_31
7. Su, Y.-S., Wu, S.-Y.: Applying data mining techniques to explore user behaviors and watching video patterns in converged it environments. J. Ambient Intell. Humaniz. Comput. 1–8 (2021). https://doi.org/10.1007/s12652-020-02712-6
8. Shoeibi, N.: Analysis of self-presentation and self-verification of the users on Twitter. In: Rodráguez González S., et al. (eds.) DCAI 2020. AISC, vol. 1242, pp. 221–226, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-53829-3_25
9. Marmo, R.: Social media mining. In: Encyclopedia of Organizational Knowledge, Administration, and Technology, pp. 2153–2165. IGI Global (2021)
10. Balakrishnan, V., Khan, S., Arabnia, H.R.: Improving cyberbullying detection using Twitter users' psychological features and machine learning. Comput. Secur. **90**, 101710 (2020)
11. Vyawahare, M., Chatterjee, M.: Taxonomy of cyberbullying detection and prediction techniques in online social networks. In: Jain, L.C., Tsihrintzis, G.A., Balas, V.E., Sharma, D.K. (eds.) Data Communication and Networks. AISC, vol. 1049, pp. 21–37. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-0132-6_3
12. Chamoso, P., González-Briones, A., De La Prieta, F., Venyagamoorthy, G.K., Corchado, J.M.: Smart city as a distributed platform: toward a system for citizen-oriented management. Comput. Commun. **152**, 323–332 (2020)
13. Yigitcanlar, T., Butler, L., Windle, E., Desouza, K.C., Mehmood, R., Corchado, J.M.: Can building "artificially intelligent cities" safeguard humanity from natural disasters, pandemics, and other catastrophes? An urban scholar's perspective. Sensors **20**(10), 2988 (2020)
14. Casado-Vara, R., Rey, A.M.-d., Affes, S., Prieto, J., Corchado, J.M.: IoT network slicing on virtual layers of homogeneous data for improved algorithm operation in smart buildings. Future Gener. Comput. Syst. **102**, 965–977 (2020)
15. González Bedia, M., Corchado Rodríguez, J.M., et al.: A planning strategy based on variational calculus for deliberative agents (2002)

16. Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: Towards cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Comput. **24**(15), 11059–11070 (2020)
17. Gencoglu, O.: Cyberbullying detection with fairness constraints. IEEE Internet Comput. **25**, 20–29 (2020)
18. Muneer, A., Fati, S.M.: A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. Future Internet **12**(11), 187 (2020)
19. Cheng, L., Shu, K., Wu, S., Silva, Y.N., Hall, D.L., Liu, H.: Unsupervised cyberbullying detection via time-informed Gaussian mixture model. arXiv preprint arXiv:2008.02642 (2020)
20. Abbass, Z., Ali, Z., Ali, M., Akbar, B., Saleem, A.: A framework to predict social crime through Twitter tweets by using machine learning. In: 2020 IEEE 14th International Conference on Semantic Computing (ICSC), pp. 363–368 (2020)
21. Balakrishnan, V., Khan, S., Fernandez, T., Arabnia, H.R.: Cyberbullying detection on Twitter using big five and dark triad features. Pers. Individ. Differ. **141**, 252–257 (2019)
22. Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G.S., On, B.-W.: Aggression detection through deep neural model on twitter. Future Gener. Comput. Syst. **114**, 120–129 (2021)
23. Corchado, J.M., Chamoso, P., Hernández, G., Gutierrez, A.S.R., Camacho, A.R., González-Briones, A., Pinto-Santos, F., Goyenechea, E., Garcia-Retuerta, D., Alonso-Miguel, M., et al.: Deepint. net: A rapid deployment platform for smart territories. Sensors **21**(1), 236 (2021). Multidisciplinary Digital Publishing Institute