Classical Machine Learning vs Deep Learning for Detecting Cyber-Violence in Social Media

Randa Zarnoufi[1](✉) and Mounia Abik[2]

[1] FSR, Mohammed V University in Rabat, Rabat, Morocco
randa_zarnoufi@um5.ac.ma
[2] ENSIAS, Mohammed V University in Rabat, Rabat, Morocco
mounia.abik@ensias.um5.ac.ma

Abstract. Cyber-violence is a largely addressed problem in e-health researches, its focus is the detection of harmful behavior from the online user-generated text in order to prevent and protect victims. In this work, we tackle the problem of Social Media (SM) text analysis to detect the harmful content that is the common characteristic of cyber-violence acts. For that, we use classical Machine Learning (ML) based on user psychological features that we compare with Deep Learning (DL) techniques in a small dataset setting. The results were in favor of classical ML. The findings highlight that psychological characteristics extracted from user-generated text are strong predictors of his harmful behavior.

Keywords: Social Media · Cyber-violence · Harmful behavior · Classical Machine Learning · Features engineering · Deep Learning

1 Introduction

Cyber-violence can be defined as: online abuse against an individual or group, often with disruptive effects on the victims. Cyber-violence has been widely discussed in the literature under different names such as cyberstalking, hate speech, and offensive, aggressive or toxic language. However, its interest remains the detection of violent contents to protect other users. Therefore, in the present study, we consider 'cyber-violence' as any act reflecting virtual violence.

For the techniques employed in the detection of cyber violence, and after a review of the literature, we have extracted the following conclusions:

- The most of related studies in the computational field have been mainly focused on supervised ML techniques based on features often of a technical nature (e.g., key-words, user and network information).
- The DL techniques, that have been used for performance improvement of existing systems without features engineering, require large amounts of annotated data.
- The previous studies have neglected important factors for detecting violent behavior, such as, personality and human behavioral characteristics [1].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2021, CCIS 1577, pp. 223–235, 2022.
https://doi.org/10.1007/978-3-031-04447-2_15

– Finally, psychological studies related to cyber-violence recommend studying the psychological characteristics of the perpetrators' personality [2, 3].

Following these recommendations, in this work, we study the impact of user' emotions and Big Five personality traits in detecting harmful content from its written text.

Our principal motivation behind extracting violent traits from individuals' writings comes from the strong relationship between language, personality and behavior. Indeed, a wide range of studies have been established on the correlation between language use and psychological traits. Some studies show that word use differs between individuals, but *correlates with their personalities and behaviors* [4–7]. Other studies confirm the *strong relationship between SM users' writing and personality traits* [8, 9].

Our approach has several advantages, first, the detection of the harmful behaviour is based on the language of emotions and personality traits that are present in almost all types of violence through the expressions written by cyber perpetrators. Second, the harmful behavior is considered to be the common characteristic of the most forms of cyber-violence. This means that our approach is generic and scalable to other forms of cyber violence. Finally, since it is based on classical ML it can work on small dataset for which DL cannot give good performance as we will prove in this paper.

The paper is organized as follows; we first present some related works to cyber-violence detection. Later we introduce our approach for harmful behavior detection based on psychological features with classical ML that we compare with DL architectures. Finally, we present and discuss the obtained results.

2 Related Works

Even if that cyber-violence includes several forms, the most covered one by the previous studies is cyber-bullying. Cyber-bullying is defined as an aggressive and repetitive act, however, by analyzing its related studies, we found that the majority of them were concerned about the harmful behavior of this act and ignored its repetitious nature [10]. It can, therefore be considered similar to other forms of violence (e.g. hate speech, offensive language…). Consequently, for the detection of the other forms of cyber-violence we can still use the same approaches as in these studies. In our case, this harmful content is considered as a sign of user behavior that will help in cyber-violence perpetrator detection.

In general, the detection task is carried out either through classical ML techniques or DL ones. ML techniques require the engineering of features and an algorithm that performs the detection. For DL, since the features are created in an autonomous way, the detection is only based on the used algorithm that requires large amounts of annotated data to ensure good performance. Here after, we present the previous works related to these techniques.

2.1 Classical ML

In classical ML, the detection process is based on two steps; the first is *features extraction* and the second is *learning* the ML model based on the extracted features.

Features Extraction
This step relies to human engineered features that aims to find the learning criterions, which here are the elements of a harmful content. According to the survey made by [10], four main categories are used; content (e.g. abusive/profane words), psychological characters (sentiments, emotions, personality traits), user (e.g. gender and age) and network (e.g. number of followers-following, the number of Likes) based features. For instance, in this work [11] the authors used Big Five and dark triad personality features in addition to network features. In our previous work [12] Big Five traits were employed effectively in harmful content detection. [13] used the emotional states of the victims after a cyber-violence episode. User's emotions were also used in our previous work [14]. In [15] they employed user, content, activity and network features to detect cyberbullying behavior. Also [16] extracted user, text, and network-based features.

ML Algorithms
Supervised learning is the most used technique for cyber-violence detection [10]. Among the used algorithms Support Vector Machine (SVM) classifier is the most used one, for instance, in [13, 17–19]. Whereas, other techniques are also used, in [11, 16] they used Random Forest classifier and they reached good performances in cyber-bullying detection. For the same purpose Al-garadi et al. [15] trained a Random Forest classifier in addition to LibSVM, the latter was the best performing model. Logistic regression was also used in many studies [20, 21] and it shows good performance.

2.2 DL Techniques

DL has been used significantly in recent years in cyber violence detection. In [22], they have addressed the problem of hate speech detection by applying different DL architectures, namely, CNN and LSTM that was the best performing one. Tommasel et al. [23] presented an approach for automatic aggression detection based on combining SVM and DL models. Their results show that aggression detection is a rather complex task, especially when it is expressed implicitly in the text (as in irony and sarcasm).

Transfer learning was also adopted in this task. Agrawal and Awekar [24] have tested transfer learning to investigate whether the knowledge gained from DL models (CNN, LSTM, BLSTM, and BLSTM with attention) on one dataset can be used to improve the performance of cyber-bullying detection on other datasets extracted from different SM platforms. In a similar study, Dadvar and Eckert [25] have replicated the same techniques by performing a transfer learning from Twitter to a YouTube dataset showing an increase in performance.

Recently, contextual embedding with BERT was used in a multilingual context to detect offensive language [26], misogyny and aggressiveness [27] while achieving very good performance.

In summary, supervised learning is the leading approach in cyber violence studies. DL techniques remain the most powerful, but require large annotated corpora. In a small dataset setting, we think that classical ML will be the right choice. However, these techniques need a careful features engineering. We have noticed that previous studies have focused on technical features and have not considered the users' psychological factors. Although, we believe that these factors can be very useful in the detection process.

Therefore, in this work, we will explore the relationship between the online user's emotions and Big Five personality traits and its harmful behavior to show their impact on cyber-violence detection. The details of our approach are given here after.

3 Approach and Method

In our approach, we assume that the harmful behaviour of the cyber-perpetrators can be identified from their emotions and personality traits.

To test this assumption, we have adopted the supervised learning approach with classical ML techniques that go through a feature extraction step followed by a learning step. First, we have extracted the features related to user' emotional states. On these features, we have trained Ensemble ML algorithms to predict the presence or absence of user's violent content. Second, we have applied the same process with the features based on the Big Five personality traits. After that, we have combined these two types of features. Finally, we have compared the performance of the generated models with those of DL based on CNN, RNN and transformer models architectures. The objective is to prove that classical ML are more convenient in small dataset setting than DL, which will allow us to save both time and computational efforts.

As use case, we have applied our approach on the detection of cyber-harassment, which is a common form of cyber violence.

We mention that, even if our approach deals with each tweet independently; however, if we can collect a set of tweets generated by the same user, we can get a clear overview of his online behaviour.

3.1 Features Extraction

In our approach, to extract linguistic features, we adopt the open vocabulary approach [9] and [20] rather than the use of special lexicon like Linguistic Inquiry and Word Count (LIWC) [28]. The main advantage of open vocabularies is that linguistic features are automatically identified and extracted from texts written by the users themselves. Special lexicons, on the other hand, are limited to predefined word lists, therefore they cannot largely cover the words used in different types of self-expressions.

For this purpose, we have used two types of features: based on emotions lexicon and based on Big Five personality traits. To extract these features from the dataset, in addition to lexical matching we use semantic similarity with word embedding to better contextualize the matching process between words from lexicons with posts' words from the dataset. For each post word, we calculate the cosine similarity between the word vectors of that word and each word from lexicon. the effectiveness of semantic similarity has been proven in our previous work [14].

Emotions. For features we have used EmoLex [29], a lexicon extracted from tweets containing words related to the eight basic emotions proposed by Plutchik [30]: anticipation, anger, fear, confidence, surprise, sadness, joy, and disgust.

Big Five Personality Traits. As features, we have chosen the Big Five personality facets (Agreeableness, Conscientiousness, Extraversion, Neuroticism and Openness. To enlarge the coverage of this lexicon we have applied a reinforcement technique based on semantic similarity using word embedding.

3.2 ML Algorithms

After features extraction, we have trained supervised learning models to predict the presence or not of harmful content. The prediction task is a binary classification.

Since the dataset used for our implementation is of a limited size, therefore, classical ML is the most convenient. Furthermore, the dataset suffers from imbalanced classes distribution with 86% for negative class and 14% for positive one. This imbalance will create a bias in the model's decision function in favor to the majority class during the learning step, and consequently it will induce errors during the prediction step. To solve this problem, we have chosen Ensemble classifiers based on decision trees which are well known for their ability to handle imbalanced data. The idea behind Ensemble ML is that by combining weak learning models, we can produce a strong prediction model and thus improve the overall result. Namely, we have used Random Forest, Gradient Boosting, XGBoost and Adaboost, their performance will be proved in the evaluation section.

4 Evaluation

The goal of the evaluation is validating the efficiency of emotion-based and Big Five-based features in comparison with DL techniques. Further details will be presented in this section, but first we will present the resources on which we have applied our classifiers.

4.1 Materials

The used materials in our experiments are Lexicons and Dataset. The used lexicons in features extraction step are of two types: the first is related to Plutchik eight emotions and the second is related to Big Five personality traits.

Emotion's Lexicons. We have compiled these features from EmoLex[1] or NRC Sentiment and Emotion Lexicons with size of 17k unigram weighted words (see example in Table 1). NRC tool contains nine lexicons types that represent the relationship between words/phrases and the eight emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust.

[1] https://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

Table 1. Examples of emotions lexicon

Emotion	Word	Weight
Anticipation	#expecting	2.237478095
Anger	jerk	0.593667390
Fear	security	0.518031195
Trust	admitting	1.485154665
Surprise	tricks	0.936144418
Sadness	hibernate	1.067902590
Joy	yey	1.747070367
Disgust	#vomit	1.518608679

Table 2. Example of agreeableness lexicon

Word	Weight
Amazing	0.056682
A great	0.056981
Fuck	−0.120624
Fucking	−0.113133

Big Five Traits' Lexicons. The second features are based on Big Five personality traits (Agreeableness, Conscientiousness, Extraversion, Neuroticism and Openness). For that, we have used the lexicon (see Table 2 for an example of agreeableness lexicon) elaborated in the work of Schwartz[2] [9]. The original lexicon is composed of 200 entries for each Big Five trait. As we mentioned earlier this lexicon was reinforced to extend its coverage (from 1000 entries to 10000 entries in total) and hence improve the models performances as proved in our previous work [12].

Dataset. We have applied our solution on a cyber-violence dataset[3] dedicated to online harassment detection in twitter posts [31], it contains 25,000 annotated tweets labeled with "Harassing" or "Not harassing". This dataset captures five different types of harassment content: sexual, racial, appearance-related, intellectual, and political. We have decided to use the "racial" dataset, one of the most common forms of online violence, it is composed of 5000 tweets. Table 3 shows an example of two entries from this dataset.

Since this dataset is provided in raw form, we first performed a preprocessing to exclude non-meaningful elements such as URLs, stop words, @ mentions and digits.

[2] https://wwbp.org/data.html.
[3] https://github.com/Mrezvan94/Harassment-Corpus.

Table 3. Racial dataset examples

Decision	Tweet
Harassing	@asadowaisi his father forgot to board train to lahore in 1947 and left this paki pig in india
Not harassing	@brandonlee161 paki haha i'm joking how are you mate?

4.2 Experiments

Classical ML

The conducted experiments aim to explore the performance of the different ML techniques first with emotions lexicon and Big Five personality traits as features independently, and second with their combination.

After features extraction, we have run each of the four classifiers on the training dataset that we split into 80% for training and 20% for test.

We have conducted five experiments each with a ML algorithm: Random Forest, Gradient Boosting, XGBoost and AdaBoost. In addition to that, we have run Penalized SVM as baseline, this classifier is considered as a very good variant of SVM and can handle imbalanced data more accurately. Moreover, it is the most used algorithm in cyberbullying detection. These experiments are as follows:

1. Penalized SVM with class_weight = 'balanced'.
2. Random Forest classifier with 100 as the maximum number of estimators (the number of trees in the forest).
3. Gradient Boosting classifier with 100 estimators.
4. XGBoost with its basic parameters without any adjustment, except the number of estimators which we fixed at 200.
5. AdaBoost with Random Forest as the base estimator, and 100 as the maximum number of estimators.

We note that all our ML methods were implemented using Scikit-Learn[4] library.

DL Techniques

To further evaluate our models, we have conducted several experiments with the DL architectures: CNN [32], RNN-BLSTM [33] and fine-tuned transformer models like BERT [34]. Next, we will give the details of each of these experiments.

CNN. CNN network is used in many NLP tasks such as text classification while showing a good performance. In our case the built model contains:

– Input layer: a 300-dimensional embedding layer.

[4] Scikit-Learn is an open-source python machine learning library.

– Hidden layers: is a CNN (1D) with 128 convolution kernels, followed by a second Conv1D layer of 64 kernels, then a pooling layer, all separated by dropout rates of 0.3.
– Output layer: a dense layer composed of a single unit, it uses a 'sigmoid' activation function to provide probability values between 0 and 1. The closer these values are to 0, the more non-violent the content of the tweet is, and the closer these values are to 1, the more violent the content is.

RNN-BLSTM. BLSTM is a variant of LSTM working in two directions. Their advantage is that they can capture patterns, perhaps omitted by the unidirectional network, and thus build more meaningful text representations. Our BLSTM model is composed of:

– Input layer: a 300-dimensional embedding layer.
– Hidden layers: two BLSTM layers of 128 and 64 units respectively, separated by a dropout of 0.3.
– Output layer: a dense layer to recover the results with a single unit and a sigmoid activation function.

We have also tested a hybridization of CNN and BLSTM networks. We have connected a Conv1D layer of 128 units with a BLSTM type GRU layer of 64 units whose output is fed into a pooling layer.

Transformer Models. Are the latest language models that have surpassed all performance records in several NLP tasks. The most known one is BERT. Their success is primarily due to their bidirectional encoder that considers the context before and after the word. Secondly, their architecture allows parallel processing of input sequences, which results in a huge gain in computation time. These models have been used as a transfer learning in text classification with fine tuning for adaptation to specific tasks since they have been pre-trained on a general domain.

In this experiment we have used the BERT-base-uncased model (12 layers of encoders, 768-hidden, 12 attention heads, and 110M parameters), in addition to RoBERTa-base (12-layer, 768-hidden, 12-heads, 125M parameters) that is an optimized version of BERT and finally, Twitter RoBERTa fine-tuned for offensive language detection.

In our test, the models fine tuning is performed as following:

– First, the input text is pre-processed to generate the tokens and attention mask identifiers required by these models.
– Then, each model is combined with a classifier, in our case it is composed of a dropout followed by a dense layer.

All over these models, we have used the optimization function 'adam', and the loss function 'binary_crossentropy', since we target a binary classification. The network was trained on 10 epochs with a batch size equal to 100. The implementations were done using TensorFlow library, especially the Keras API. For transformer models we have

used the *transformers* library developed by Hugging Face[5].We mention that we have split the racial dataset into 70% for training, 20% for validation and 10% for testing.

4.3 Results and Discussion

Evaluation Metrics. As we said before, our dataset is imbalanced (86% for negative class and 14% for positive one). Consequently, a classifier that does not take into consideration the imbalanced class issue will generate an overfitting by only predicting the majority class with a high accuracy. In such situation, Accuracy is no longer a suitable metric. This is why we have chosen the AUC (Area Under the Curve ROC) associated with the ROC (Receiver Operating Characteristic) curve as the main metric:

$$AUC = \frac{1 + TP_{Rate} - FP_{Rate}}{2} \tag{1}$$

AUC (formula 1) is widely used as an evaluation metric in case of imbalanced class distribution. The ROC curve plots the true positive rate (TP_{rate}) against the false positive rate (FP_{rate}), allowing the separation of signal (TP) from noise (FP). The AUC is the area under the ROC curve and is considered as a summary of the ROC curve. The AUC measures the ability of a model to differentiate between classes. The larger the AUC value, the better the model is at differentiating between positive and negative classes.

Results. The results obtained from the different experiments are illustrated in the tables below. We note that all metrics are given in macro-average.

Classical ML Results

a. *Emotions-based learning results*
Table 4 illustrate the results given by the five classifiers, as it is shown, XGBoost has achieved the best results in terms of AUC 0.75. The second to best classifier was Gradient Boosting with an AUC score of 0.73, then Adaboost with 0.72. Among the five classifiers, the penalized SVM scored the lowest in all metrics.

Table 4. Classifiers' performance results with Emotions as features

Classifier	AUC	Precision	Recall	F1
Penalized SVM	0.53	0.42	0.50	0.45
Random forest	0.71	0.78	0.58	0.66
gradient boosting	0.73	0.73	0.54	0.62
XGBoost	0.75	0.74	0.55	0.63
AdaBoost	0.72	0.71	0.59	0.64

b. *Big Five-based learning results*
Table 5 shows the results obtained from the experiments where we compare the performance of the five classifiers (with lexicon reinforcement).

[5] https://huggingface.co.

Table 5. Classifiers' performance results with Big Five as features

Classifier	AUC	Precision	Recall	F1
Penalized SVM	0.5	0.79	0.54	0.64
Random Forest	0.73	0.77	0.64	0.69
Gradient Boosting	0.71	0.75	0.52	0.61
XGBoost	0.72	0.65	0.65	0.65
AdaBoost	0.72	0.82	0.61	0.69

As shown in this Table 5, the best AUC score was achieved by Random Forest (0.73). AdaBoost achieved the best results in terms of precision and F1 (0.82, 0.63 respectively). XGBoost reached the highest recall among all other classifiers (0.65). Finally, among the five classifiers, penalized SVM performed the lowest in all metrics except precision (0.79) where SVM was ranked second.

c. *Emotions and Big Five traits Combination*
To evaluate the impact of emotions and Big Five features combination on this task, we have conducted this third experiment. The results are presented in Table 6 showing an increase in performance especially in AUC that has reached 0.80, which means that the combination of personality features was more efficient in this task.

Table 6. Results of emotions and Big Five features combination

Classifier	AUC	Precision	Recall	F1
Penalized SVM	0.38	0.13	0.50	0.20
Random Forest	0.73	0.70	0.54	0.60
Gradient Boosting	0.79	0.76	0.57	0.65
XGBoost	0.80	0.77	0.59	0.66
AdaBoost	0.74	0.79	0.56	0.65

Next, we give the results of harmful content detection with DL techniques.

DL Results

Table 7 presents the results given by the DL architectures CNN, BLSTM in addition to transformer models: BERT, RoBERTa and Twitter RoBERTa for offensive language detection. As observed, all DL models show poor performance over all metrics except for RoBERTa fine-tuned on offensive language which has achieved quite good results.

In summary, ensemble ML techniques have proved their performance for the case of small and imbalanced dataset. Although DL techniques are known for their high performance in many NLP tasks, however, they require large amounts of data to achieve such performance. This was confirmed by the low scores of different evaluation metrics.

Table 7. Performance of DL architectures CNN, BLSTM and transformer models

Classifier	AUC	Precision	Recall	F1
CNN	0.53	0.43	0.50	0.46
BLTSM	0.49	0.43	0.50	0.46
CNN + BLSTM (GRU)	0.42	0.43	0.50	0.46
BERT	0.57	0.58	0.62	0.59
RoBERTa	–	0.43	0.50	0.46
Twitter RoBERTa Offensive	0.67	0.67	0.77	0.71

In contrast to classical ML techniques which can achieve good results even with a small dataset. Regarding transformer models, as they were trained on a very large amount of general domain corpora, they need to be fine-tuned on specific domain to provide better results, which was proven by the good recall reached by RoBERTa for offensive language model. Finally, these findings show that user psychological characteristics extracted from its written text can be good indicators of its online harmful behavior.

5 Conclusion

To help in individuals' well-being, we are interested in this study in finding a mean to automatic detection of harmful behavior from the online users' generated text. Which can lead to the detection of cyber-perpetrators.

Psychologists state that cyber-violence act is related to the perpetrator's psychology. Along this study, we tried to demonstrate the validity of this assumption, where, we extracted features related to personality and we trained supervised models on racial harassment dataset. In particular, we used Ensemble Machine Learning that have shown good performance in dealing with imbalanced dataset.

We have also proved that classical ML can outperform DL techniques in a small dataset context while saving computational efforts. However, transfer learning with transformer models is still appealing in case of further fine-tuning with specific dataset.

The obtained results show that individual's psychological features are correlated with his/her harmful behavior. Furthermore, our solution can be generalized to be employed in detecting other type of cyber-violence where harmful behaviors are present as in hate speech for instance. Finally, these findings may be exploited in e-health interventions by the organizations interested to this phenomenon.

References

1. Sanchez, H., Kumar, S.: Twitter bullying detection. In: NSDI, pp. 15–22 (2011)
2. Kowalski, R.M., Giumetti, G.W., Schroeder, A.N., Lattanner, M.R.: Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Psychol. Bull. © 2014 Am. Psychol. Assoc. **140**, 1073–1137 (2014)

3. Paul, S., Smith, P.K., Blumberg, H.H.: Investigating legal aspects of cyberbullying. Psicothema **24**, 640–645 (2012)
4. Davahli, M.R., et al.: Personality and text: quantitative psycholinguistic analysis of a stylistically differentiated Czech text. Psychol. Stud. (Mysore). **12**, 1–23 (2020)
5. Moreno, J.D., Martínez-Huertas, J., Olmos, R., Jorge-Botana, G., Botella, J.: Can personality traits be measured analyzing written language? A meta-analytic study on computational methods. Pers. Individ. Dif. **177** (2021)
6. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**, 24–54 (2010)
7. Yarkoni, T.: Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers. J. Res. Pers. **44**, 363–373 (2010)
8. Azucar, D., Marengo, D., Settanni, M.: Predicting the big 5 personality traits from digital footprints on social media: a meta-analysis. Pers. Individ. Dif. **124**, 150–159 (2018)
9. Schwartz, H.A., et al.: Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS ONE **8**, e73791 (2013)
10. Salawu, S., He, Y., Lumsden, J.: Approaches to automated detection of cyberbullying: a survey. IEEE Trans. Affect. Comput. **3045**, 1–20 (2017)
11. Balakrishnan, V., Khan, S., Fernandez, T., Arabnia, H.R.: Cyberbullying detection on twitter using big five and dark triad features. Pers. Individ. Dif. **141**, 252–257 (2019)
12. Zarnoufi, R., Abik, M.: Big five personality traits and ensemble machine learning to detect cyber-violence in social media. In: Serrhini, M., Silva, C., Aljahdali, S. (eds.) EMENA-ISTL 2019. LAIS, vol. 7, pp. 194–202. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-36778-7_21
13. Dadvar, M., Ordelman, R., de Jong, F., Trieschnigg, D.: Towards user modelling in the combat against cyberbullying. In: Bouma, G., Ittoo, A., Métais, E., Wortmann, H. (eds.) NLDB 2012. LNCS, vol. 7337, pp. 277–283. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31178-9_34
14. Zarnoufi, R., Boutbi, M., Abik, M.: AI to prevent cyber-violence: harmful behaviour detection in social media. Int. J. High Perform. Syst. Arch. **9**, 182–191 (2020)
15. Algaradi, M.A., Varathan, K.D., Ravana, S.D.: Computers in human behavior cybercrime detection in online communications: the experimental case of cyberbullying detection in the Twitter network. Comput. Human Behav. **63**, 433–443 (2016)
16. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A.: Mean birds: detecting aggression and bullying on Twitter. In: Proceedings of the 2017 ACM on Web Science Conference, New York, USA, pp. 13–22 (2017)
17. Dadvar, M., de Jong, F., Ordelman, R., Trieschnigg, D.: Improved cyberbullying detection using gender information. In: 12th - Dutch-Belgian Information Retrieval Workshop. DIR'2012, pp. 22–25 (2012)
18. Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q., Mishra, S.: Detection of cyberbullying incidents on the Instagram social network. In: 13th Annual International Conference on Mobile Systems, Applications, and Services, Florence, 18–22 May 2015, p. 481. ACM (2015)
19. Robinson, D., Zhang, Z., Tepper, J.: Hate speech detection on Twitter: feature engineering v.s. feature selection. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 11155, pp. 46–49. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98192-5_9
20. Stillwell, D., Matz, S.: Latent human traits in the language of social media: an open-vocabulary approach latent human traits in the language of social media. PLoS ONE **13**(11) (2018)
21. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of NAACL-HLT, pp. 88–93 (2016)

22. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759–760 (2017)
23. Tommasel, A., Rodriguez, J.M., Godoy, D.: Textual aggression detection through deep learning. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pp. 177–187 (2018)
24. Agrawal, S., Awekar, A.: Deep learning for detecting cyberbullying across multiple social media platforms. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 141–153. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_11
25. Dadvar, M., Eckert, K.: Cyberbullying detection in social networks using deep learning based models; a reproducibility study. In: DaWaK, pp. 1–13 (2018)
26. Ranasinghe, T., Zampieri, M., Hettiarachchi, H.: BRUMS at HASOC 2019: deep learning models for multilingual hate speech and offensive language identification. In: FIRE 2019 (2019)
27. Samghabadi, N.S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., Solorio, T.: Aggression and misogyny detection using BERT: a multi-task approach. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying LREC 2020, pp. 126–131 (2020)
28. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015 (2015)
29. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Comput. Intell. **29**, 436–465 (2013)
30. Plutchik, R.: Emotion: a psychoevolutionary synthesis (1980)
31. Rezvan, M., Shalin, V.L., Sheth, A.: A quality type-aware annotated corpus and lexicon for harassment research. In: WebSci 2018, Web Science. ACM (2018)
32. Lecun, Y., et al.: Handwritten digit recognition with a back-propagation network. In: NIPS, pp. 396–404 (1990)
33. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Sig. Process. **45**, 2673–2681 (1997)
34. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, pp. 4171–4186 (2019)