Cyber Aggression and Cyberbullying Identification on Social Networks

Vincenzo Gattulli[1], Donato Impedovo[2], Giuseppe Pirlo[2] and Lucia Sarcinella[2]
[1]*Digital Innovation Srl, Via Edoardo Orabona, 4 (c/o Dipartimento di Informatica), 70125 Bari, Italy*
[2]*Department of Computer Science, University of Studies of Bari "Aldo Moro", Via Edoardo Orabona, 4, 70125 Bari, Italy*

Keywords: Cyberbullying, Artificial Intelligence, Social Network, Cyber Aggression, Twitter, Machine Learning.

Abstract: Bullying includes aggression, harassment, and discrimination. The phenomenon has widespread with the great diffusion of many social networks. Thus, the cyber aggression iteration turns into a more serious problem called Cyberbullying. In this work an automatic identification system built up on the most performing set of techniques available in literature is presented. Textual comments of various Italian Twitter posts have been processed to identify the aggressive phenomenon. The challenge has been also identifying aggressive profiles who repeat their malicious work on social networks. Two different experiments have been performed with the aim of the detection of Cyber Aggression and Cyberbullying. The best results were obtained by the Random Forest classifier, trained on an ad-hoc Dataset that contemplates a series of comments extracted from Twitter and tagged manually. The system currently presented is an excellent tool to counter the phenomenon of Cyberbullying, but there are certainly many improvements to be made to improve the performance of the system.

1 INTRODUCTION

Social Networks are progressively exposed to harmful issues including Cyber Aggression and Cyberbullying. Cyber Aggression refers to aggressive online behaviour using digital media content *(text, images, videos, etc.)* to cause harm to another person. Cyberbullying is defined as "An aggressive intentional act by an individual or a group of individuals, using electronic forms of contact, repeated over time against a victim who cannot easily defend himself" (Dredge et al., 2014).

This work deals with the automatic recognition of Cyber Aggression (detected in textual comments) and cyberbully profiling (Cyber Aggression repeated by a certain user in multiple posts over time). The experiments aim to prevent the phenomenon of Cyberbullying on Social Networks. In this study Italian behavioral patterns will be studied, recovering recurrent patterns in the formulation of Italian sentences or in the typology of attacks with the aim of collecting and labeling comments and creating a dataset called "*Italian Aggressive Dataset*". The attention is mainly focused on the kind of language used by the attacker considering vocabularies of expressions and words belonging to the vulgar jargon. Next, each word within the sentence is weighted according to its negative, neutral, or positive value along with a large set of other characteristics. The main contributions of this work are:

- An ensemble of features with a comparison of different classification models.
- The creation of a vocabulary of Italian words considering four types of categories: *Bad Word, Second Person, Threats, Bulling Terms. These dictionaries contain some of the most common terms in Italian, used to verbally attack and offend someone. The definition of these words was made by viewing countless comments under the posts of famous singers and politicians.*
- "*Aggressive Italian Dataset*": Creation and labeling of a balanced Italian dataset composed of aggressive and non-aggressive comments, extracted from the social platform Twitter named.

The rest of the work is organized as follows. Section 2 will illustrate the state of the art. Section 3 will describe the software design and implementation methodology. Section 4 will describe the "*Italian Aggressive Dataset*". The results of the experimentation are provided in Section 5. Finally, Section 6 concludes the document.

Gattulli, V., Impedovo, D., Pirlo, G. and Sarcinella, L.
Cyber Aggression and Cyberbullying Identification on Social Networks.
DOI: 10.5220/0010877600003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 644-651
ISBN: 978-989-758-549-4; ISSN: 2184-4313

2 RELATED WORK

This document focuses on textual data being the most widespread data in social media having the aim in identifying aggression and pattern could be referred to cyberbullies in their mild stage (Shah et al., 2021). Many researchers have worked on textual comments collected on Social Networks. Amali et. al (Ishara Amali & Jayalal, 2020) with the aim of determining insults (profane words) in the comments within the tweets, five rules were taken into consideration (Ishara Amali & Jayalal, 2020): i) percentage of bad words within the tweet, ii) combination of first-person pronoun, bad word and a second person pronoun, iii) combination of second person pronoun with a bad word combination of third person pronoun with a bad word, iv) combination of first-person pronoun, bad word and a third person pronoun. Selected comments were successively classified adopting SVM, K-Nearest Neighbors (KNN) and Naïve Bayes (NB): the SVM with RBF kernels scored better than others reporting a 91% f1-score.

Chatzakou et al. (Chatzakou et al., 2017) processed 1.6 million tweets collected over 3 months of conversations. In this case user-based features, text-based features and network-based features were considered. User-based features had the aim of describing the general user's behaviour (e.g., bully, and generic aggressors are faster than normal users in posting activity), text-based features were referred to uppercases, specific word embedding and to the positive/negative sentiment in short text. The network-based features were aimed to evaluate popularity, reciprocity, power difference and influence of users within the group. RF classifier was able to perform 90% of AUC (Chatzakou et al., 2017). Raza et al. (Raza et al., 2020) developed a model with LR, RF and NB algorithms to identify if a particular comment is an insult, threat, or a hate message, with Voting and AdaBoost classifiers. Supervised machine learning with LR achieved 82.7% accuracy. With the voting classifier, an accuracy of 84.4% was observed (Raza et al., 2020). Shtovba et al. (Shtovba et al., 2019) found syntactic dependencies in comments, i.e., relationships with proper nouns, personal pronouns, possessive pronouns, etc. Three features were highlighted that greatly improve detection: the number of dependencies with proper names in the singular, the number of dependencies that contain profanity, and the number of dependencies between personal pronouns and profanities. The data used comes from the Kaggle contest "Toxic Comment Classification Challenge (Large number of Wikipedia comments)". An DT classifier is used (Shtovba et al., 2019).

Dwivedi et al. (Kumari et al., 2021b) present a deep learning-based model (LSTM network) detecting different levels of aggression (direct, indirect and no aggression) in social media posts in a bilingual scenario. Datasets from Facebook and Twitter with bilingual (English and Hindi) data were used (Kumari et al., 2021b). Sentiment description has been also considered evaluating comments i) contain remarks, critic, sarcasm, etc., ii) referred to specific topics (e.g., politics, crimes, race, sex, etc.), iii) containing swear words. In this case three different classifiers were adopted: KNN, SVM, and LR. The best performances were achieved by the SVM with linear kernel reporting 86% on accuracy and recall and 84% of f1 score (Chen et al., 2017). The automatic detection of cyberbullying can be exploited considering psychological features of users, including personalities, feelings, and emotions. User personalities can be determined using the Big Five model (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) (Costa & McCrae, 1992) (John & Srivastava, n.d.) and Dark Triad (narcissism, Machiavellianism, and psychopathy) which specifically refers to malevolent qualities (Paulhus & Williams, 2002) (Goodboy & Martin, 2015). Many scientific papers have used hybrid approaches, correlating images (Dentamaro et al., 2021) and post comments under the same images. Singh et al. (Kumari et al., 2021a) present a model based on Convolutional Neural Network (CNN) and Binary Particle Swarm Optimization (BPSO) to classify social media posts containing images with associated textual comments into non-aggressive, medium aggressive and highly aggressive classes. The proposed model with optimized features and Random Forest classifier (Dentamaro et al., 2020) achieves a weighted F1-Score of 0.74 (Kumari et al., 2021a). Kumari et al. (Kumari & Singh, 2021a) present textual features extracted using a three-layer parallel convolutional neural network. The image and text features are then combined to obtain a hybrid feature set that is further optimized using a binary firefly optimization algorithm (Kumari & Singh, 2021a). Finally, Singh et al. (Kumari & Singh, 2021b) present a pre-trained VGG-16 network and a convolutional neural network to extract features from images and text, respectively. These features are further optimized using a genetic algorithm to increase the efficiency of the whole system. The proposed model achieves an F1 score of 78% (Kumari & Singh, 2021b). The hybrid approach was not considered due to both the lack of datasets and the poor performance reported in the read article.

3 METHODS

The proposed approach is organized in a pipeline made-up of three stages:

A. Post selection and test comments extraction;
B. Feature engineering;
C. Classification and Metrics.

This paper proposes two different experiments. The first one aims to identify Cyber Aggression, the second one aims to identify Cyberbullying. The first experiment identifies aggression from user comments. In case an aggression is identified by the classifier and there are multiple aggressions on multiple posts by the same user, then that user could be flagged as an aggressive profile (bully), thus giving rise to the second phase. The system is designed not only to run experiments as described in this paper, but also to be able to work online. For the Training phase, the dataset created in this study called the *Aggressive Italian Dataset* is used. For the testing phase, an additional 1000 different comments were extracted from different Twitter posts for each of the four celebrities that we will discuss in the next subchapter. The importance of identifying the different posts is related to the problem of identifying cyberbullies stalking the victim. The comments selected for the Test phase were manually labeled, and the feature extraction phase was performed for each comment, as well as for the Training. Twitter comments from the Test phase extracted are in Italian, dated November-December 2020. During the period considered, each post contained approximately 100/150 comments (6 Twitter posts). In summary, the "Aggressive Italian Dataset" containing 3028 comments was used for the Training phase. As a Test, 1000 comments of different posts were extracted for each famous person. Figure 1 illustrates the phases of the experiment.

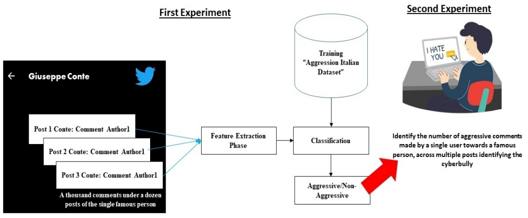
First Experiment
Second Experiment
Giuseppe Conte
Post 1 Conte: Comment Author1
Post 2 Conte: Comment Author1
Post 3 Conte: Comment Author1
A thousand comments under a dozen posts of the single famous person
Feature Extraction Phase
Training "Aggression Italian Dataset"
Classification
Aggressive/Non-Aggressive
I HATE YOU
Identify the number of aggressive comments made by a single user towards a famous person, across multiple posts identifying the cyberbully

Figure 1: General scheme of the system.

A. Post Selection and Test Comments Extraction

Famous people with large audiences and many followers clearly attract both supporters and "haters". Four famous Italian people who suffer acts of Cyber Aggression have been considered in this work. Many users carry out verbal aggression by commenting under each post, also attacking private life, behavior very similar to a Cyberbullying action. Profiles here considered are: *Achille Lauro (Italian singer/rapper), Fabio Rovazzi (Italian singer and youtuber), Matteo Renzi and Giuseppe Conte (Italian politicians).* The period selected for posts is between November-December 2020. In this period, Italy was in a government crisis and a coronavirus pandemic.

B. Feature Engineering

In the feature extraction phase, nine features are considered:

Number of Negative Words (BW). This feature has been implemented by means of a "BadWord" vocabulary containing 540 negative words extremely vulgar used for aggressive purposes, offenses, and humiliations. The use of regular expressions has allowed to identify also negative words written grammatically incorrect, all attached or with spaces (Ex. Assssshole → asshole) (Ishara Amali & Jayalal, 2020).

Number of "non/no"(NN). The use of "no/not" within a sentence completely changes the meaning of the sentence from positive to negative or vice versa. Furthermore, the presence of a large number of "no/not" can underline the controversy of the comment.

Uppercase (U). This is a Boolean value indicating whether the comment is capitalized or not. In computer jargon, uppercase comment is about shouting something. So, it can be interpreted as an aggression against someone (Chatzakou et al., 2017).

Positive/negative Weight of the Comment (PW/NW). This feature includes two values: a positive and negative weight of the comment within the range [0,1]. To do this, the relative synset and weight of each word was extracted, using WordNet and SentiWordNet (M et al., 2017) (Rendalkar & Chandankhede, 2018) and then averaged for both positive and negative weights. The average value was chosen to take into account the length of the comment and therefore the number of words.

Use of the Second Person (SP). It is a Boolean value indicating the presence or absence of a second singular or plural form in the comment. This feature is important because attacks are often accompanied using the second person, thus targeting a specific person. This feature was extracted through a specially created dictionary containing 24 words, including verbs and pronouns referring to the second person, for example words ending with "tu" or "ti" ("you" in English) (Shtovba et al., 2019).

Presence of Threats (TR), instigation to violence or suicide. A Boolean value indicates the presence of threats, violence, or instigation to suicide within the comments. Many negative comments are accompanied using profane language or threats such as "I kill you" or instigations to suicide "thrown from a bridge", all expressions used only in aggressive contexts. Even in this case, these expressions have been identified using a specifically devoted vocabulary containing 314 violent or instigating expressions (Raza et al., 2020).

Presence of Bulling Terms (KW). A Boolean value indicates the presence of the so-called keywords of cyberbullying, or insults, used to injure or attack a person (e.g., idiot, stupid, ...), but also target words which in themselves do not take on a negative meaning, but in some contexts, such as that of cyberbullying, they can be used equally to insult (e.g., clown, whale, garbage, ...). In this case a vocabulary containing 359 terms identified as insults and possible insults has been created.

Comment Length (L). This feature represents the length of the comment in terms of words. In fact, it has been observed that most negative comments are made up of a few words, usually no more than three.

The choice of these nine features was dictated by both the state of the art and a careful analysis made on the real comments of the people of Twitter. It has been found that the language on the web is rough, full of expressions and words belonging to the vulgar jargon that leaves little room for misunderstanding. The Italian language has many identical terms that are used when verbally attacking someone. This led to the creation of a list of these words, creating a veritable dictionary of profane words. The weight of profane words does not have to decree with certainty the negativity of the sentence, for this reason another feature has been devised that considers the weight that a word can have in a sentence, both in negative and in positive. Again, it was noted that some negative comments were capitalized, as if to simulate a higher tone of voice. This has led to thinking of a way to keep track of this particularity. Another feature that was highlighted is the presence of negation in aggressive comments, in fact in many cases it was noted that the aggression sessions started with the word "no / non" to contradict the victim. Again, the presence of the second person, an example would be "TI uccido" (I kill you), "DEVI morire" (You must die). As an enrichment of the vocabulary on "profane words", two other vocabularies have been defined with expressions very close to aggressive juvenile language. The first is defined as expressions of incitement to violence with the purpose of wishing someone's death. The second are defined as expressions linked to juvenile and offensive language, closely linked to pokes and assonances with animals in a derogatory way. Aggression in both bullying and in a more general context embraces these themes which have been gradually considered and applied to the extraction of each individual comment.

C. Classification and Matrics

The classification of the comments has been carried out using four supervised classification algorithms: SVM with linear kernel, RF, MLP and DT. The problem of classification has been considered here as a two class one: Aggressive comments and Non-Aggressive ones. In this work the SVM kernel is linear because it works well for text classification (Malmasi & Zampieri, 2018) (Davidson et al., n.d.). In this work the maximum RF depth has been set at 10, and the number of estimators is set at 1800 (Islam et al., 2019) (Chatzakou et al., 2017). In this work the MLP alpha parameter has been set equal to 0.05, hidden layer levels equal to (25, 20) and learning rate equal to 0,001 (Ramchoun et al., 2016). The parameters considered were tested as best after a Greed Search approach. Four parameters were considered to evaluate the system performance: *Accuracy, Precision (P), Recall (R), F1-score (F1)* (Prastowo et al., 2019).

4 DATASET

The "*Aggressive Italian Dataset*" consists of Italian comments extracted from Twitter, both Aggressive and Non-Aggressive and contains 3028 comments. Comments were divided between 1514 aggressive and 1514 non-aggressive. The dataset was carefully balanced keeping the same number of aggressive and non-aggressive comments, labelled (T) with the manual procedure explained below. Each comment was analyzed by ten people, each person categorized the comment as aggressive and non-aggressive through their attitude towards the issue.

Finally, the most frequent classes were assigned to each of the comments. Aggression was understood as any form of aggression that hurt the sensibilities of the person being attacked. The content of the comments did not have to contain a profane word, but a verbal attack that could hurt any person receiving that message. While, about comments classified as non-aggressive, those comments that did not go to hurt the sensitivity of others were considered. After labeling, statistically it was noted that the people in question agreed because the selected comments carry little ambiguity. Many aggressive comments registered feelings of violence and aggression account a particular person. If the dataset will be extended and shared the labeling part will be better specified.

Table 1: Evaluation of the comments of the last six posts by Achille Lauro.

	SVM			DT			RF			MLP		
Achille Lauro	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Not-aggressive	0.98	0.88	0.93	0.94	0.86	0.90	0.99	0.91	0.95	0.96	0.91	0.94
Aggressive	0.70	0.94	0.81	0.64	0.83	0.72	0.77	0.98	0.86	0.75	0.75	0.81
Accuracy	0.90			0.85			0.93			0.90		

Table 2: Evaluation of the comments of the six posts of Fabio Rovazzi.

	SVM			DT			RF			MLP		
Fabio Rovazzi	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Not-aggressive	0.94	0.84	0.89	0.89	0.78	0.84	0.98	0.83	0.90	0.92	0.86	0.89
Aggressive	0.75	0.90	0.82	0.66	0.82	0.73	0.75	0.97	0.85	0.76	0.87	0.81
Accuracy	0.86			0.80			0.88			0.86		

Table 3: Evaluation of the comments of the last six posts by Matteo Renzi.

	SVM			DT			RF			MLP		
Matteo Renzi	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Not-aggressive	0.98	0.95	0.97	0.98	0.95	0.96	0.99	0.98	0.98	0.98	0.96	0.97
Aggressive	0.74	0.89	0.81	0.71	0.84	0.77	0.85	0.95	0.90	0.76	0.88	0.82
Accuracy	0.94			0.94			0.97			0.95		

Table 4: Evaluation of the comments of the last six posts by Giuseppe Conte.

	SVM			DT			RF			MLP		
Giuseppe Conte	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Not-aggressive	0.93	0.84	0.89	0.89	0.77	0.83	0.96	0.85	0.90	0.90	0.85	0.88
Aggressive	0.73	0.87	0.79	0.64	0.81	0.71	0.75	0.92	0.82	0.82	0.84	0.78
Accuracy	0.85			0.80			0.87			0.84		

Table 5: Extract of the table containing the profiles of cyberbullies.

Twitter Users	nAC	nC	nAC Lauro	nAC Rovazzi	nAC Renzi	nAC Conte
Peppe***	7	12	7	0	0	0
paoloG***	6	8	0	3	3	0
nonseic***	6	8	6	0	0	0
peso***	6	8	0	6	0	0
bettav***	5	5	0	0	5	0
marcoLu**	4	9	0	0	4	0
alessand_*	4	4	0	4	0	0
fatazu****	4	13	0	0	0	4
Mart***	4	4	0	4	0	0
…	…	…	…	…	…	…

Table 6: Extract of table containing the profiles of the victims.

RF Model	Number of aggressive comments predicted	Number of actual aggressive comments	Aggressive comments classification error percentage
Achille Lauro	123	129	5%
Fabio Rovazzi	224	229	2%
Matteo Renzi	332	342	3%
Giuseppe Conte	302	329	8%

5 EXPERIMENTS

5.1 First Experiment

In the first experiments are revealed multiple observations. The results for the different characters are very similar to each other, but the RF model achieved the best results considering the Accuracy value for each of the four famous characters. In fact, the results of the RF model regarding the identification of comments without Cyber Aggression, P is between 96-99%, R between 83-98% and F1-score values between 90-98%. As for the identification of aggressive comments, the RF achieves an P ranging from 75% to 84%, R values ranging from 92% to 98%, F1 score values ranging from 82% to 90%. The results are shown in Tables 1, 2, 3 and 4. In general, however, the accuracy of the entire system is very high, with values ranging from 80% to 98%. The cases in which the accuracy assumes very low values, are within the posts in which there were several ironic comments or offensive words hidden by an incorrect use of grammar, not identified by the system. All capitalized comments that do not present aggression could be identified as non-aggressive, or situations in which the aggression does not refer to the person in the post, but to another of similar social class. Hashtags that may contain vulgar and offensive slogans are not recognized.

5.2 Second Experiment

Based on considerations previously reported, a basic and easy to implement approach to reveal a tendency or a propensity to a trait of cyberbullying is to evaluate recurrence of aggressions. To the aim the total number of comments considered as Cyber Aggression by the system have been enumerated. Two tables were created. In the first, all the users who have been identified for a Cyber Aggression through their comments are stored, and for each of them the total number of comments, the number of comments identified as Cyber Aggression and the related posts are marked. With this first table it is therefore possible to identify the cyberbully, detecting who has many cyber-attacks in general or aimed at a specific victim. An example would be the user named *"paoloG*****" who wrote 8 comments, and more than half were classified as Cyber Aggressive. In addition, three were addressed on several posts to the politician Matteo Renzi and three were addressed to the singer Fabio Rovazzi. This means that this user does not attack a single social category but attacks more than a semantically different social category. A simpler example could be the user "mart****" who made four comments in several posts by Fabio Rovazzi, all classified as Cyber Aggressive (see Table 5).

The second table (see Table 6), on the other hand, shows the results of the Random Forest. It can be seen that for all four characters, a 10% detection error was made out of 1000 standard comments in the Test phase (see Table 6). In the tables, the wording "nC" identifies the number of comments made by the particular user, while the wording "nAC" identifies the aggressive comments predicted by the system.

6 CONCLUSIONS

In this work, the problem of Cyber Aggression related to Cyberbullying was considered. The results obtained show that the identification of aggressive comments is done with a good degree of accuracy. Different classification schemes were compared, and Random Forest (RF) was found to be the one that achieved the highest accuracy for all the different cases considered here. The next step was the identification of Cyberbullying sessions by tracking users who posted comments classified as aggressive. Through this analysis it was possible to obtain the profiles of the pages most prone to this type of phenomenon, being able to monitor these victims, to report the situation to the competent authorities. In addition, in this work have been considered some innovative Feature Engineering phases and some state-of-the-art ones, that have allowed together with the creation of the "Aggressive Italian Dataset" in Italian, the possibility to identify common patterns in the Italian culture that could identify an aggression. In addition to the macro dataset, innovative sub vocabularies were created that also allowed the identification of verbal aggression. This procedure could be performed online, in a smaller context such as schools to prevent cyberbullying. These datasets could only be shared if the article is accepted. Regarding future developments, the detection of Cyberbullying comments should be improved, as it was noted that many slang forms, or even grammatically incorrect, were not identified, so it is possible to expand the vocabularies used. It is also possible to make a deeper analysis by monitoring victims and cyberbullies to understand the frequency or reasons that lead to this phenomenon and thus prevent them. Finally, it would be interesting to greatly expand the dataset by adding comments to the posts of other people, not necessarily famous, but ordinary people.

ACKNOWLEDGEMENTS

This work is supported by the Italian Ministry of Education, University and Research within the *PRIN2017 - BullyBuster project - A framework for bullying and cyberbullying action detection by computer vision and artificial intelligence methods and algorithms*.

REFERENCES

Chatzakou, D., Kourtellis, N., Blackburn, J., de Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter. WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference, 13–22. https://arxiv.org/abs/1702.06877v3

Chen, J., Yan, S., & Wong, K. C. (2017). Aggressivity detection on social network comments. ACM International Conference Proceeding Series, Part F127854, 103–107. https://doi.org/10.1145/3059336.3059348

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. Personality and Individual Differences, 13(6), 653–665. https://doi.org/10.1016/0191-8869(92)90236-I

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (n.d.). Automated Hate Speech Detection and the Problem of Offensive Language *. Retrieved September 14, 2021, from www.facebook.

Dentamaro, V., Giglio, P., Impedovo, D., & Pirlo, G. (2021). Benchmarking of Shallow Learning and Deep Learning Techniques with Transfer Learning for Neurodegenerative Disease Assessment Through Handwriting. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12917 LNCS, 7–20. https://doi.org/10.1007/978-3-030-86159-9_1

Dentamaro, V., Impedovo, D., & Pirlo, G. (2020). Gait Analysis for Early Neurodegenerative Diseases Classification through the Kinematic Theory of Rapid Human Movements. IEEE Access, 8, 193966–193980. https://doi.org/10.1109/ACCESS.2020.3032202

Dredge, R., Gleeson, J., & de La Piedad Garcia, X. (2014). Presentation on Facebook and risk of cyberbullying victimisation. Computers in Human Behavior, 40, 16–22. https://doi.org/10.1016/J.CHB.2014.07.035

Goodboy, A. K., & Martin, M. M. (2015). The personality profile of a cyberbully: Examining the Dark Triad. Computers in Human Behavior, 49, 1–4. https://doi.org/10.1016/J.CHB.2015.02.052

Ishara Amali, H. M. A., & Jayalal, S. (2020). Classification of Cyberbullying Sinhala Language Comments on Social Media. MERCon 2020 - 6th International Multidisciplinary Moratuwa Engineering Research Conference, Proceedings, 266–271. https://doi.org/10.1109/MERCON50084.2020.9185209

Islam, M. Z., Liu, J., Li, J., Liu, L., & Kang, W. (2019). A semantics aware random forest for text classification. International Conference on Information and Knowledge Management, Proceedings, 1061–1070. https://doi.org/10.1145/3357384.3357891

John, O. P., & Srivastava, S. (n.d.). The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives.

Kumari, K., & Singh, J. P. (2021a). Multi-modal cyber-aggression detection with feature optimization by firefly algorithm. Multimedia Systems. https://doi.org/10.1007/S00530-021-00785-7

Kumari, K., & Singh, J. P. (2021b). Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Transactions on Emerging Telecommunications Technologies, 32(2), e3907. https://doi.org/10.1002/ETT.3907

Kumari, K., Singh, J. P., Dwivedi, Y. K., & Rana, N. P. (2021a). Bilingual Cyber-aggression detection on social media using LSTM autoencoder. Soft Computing 2021 25:14, 25(14), 8999–9012. https://doi.org/10.1007/S00500-021-05817-Y

Kumari, K., Singh, J. P., Dwivedi, Y. K., & Rana, N. P. (2021b). Multi-modal aggression identification using Convolutional Neural Network and Binary Particle Swarm Optimization. Future Generation Computer Systems, 118, 187–197. https://doi.org/10.1016/J.FUTURE.2021.01.014

M, R. v, Kumar, M. P., Raman, S. R., & Sridhar, R. (2017). Emotion And Sarcasm Identification Of Posts From Facebook Data Using A Hybrid Approach. Online) Ictact Journal On Soft Computing, 2. https://doi.org/10.21917/ijsc.2017.0197

Malmasi, S., & Zampieri, M. (2018). Challenges in Discriminating Profanity from Hate Speech. Journal of Experimental and Theoretical Artificial Intelligence, 30(2), 187–202. https://arxiv.org/abs/1803.05495v1

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. Journal of Research in Personality, 36(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Prastowo, E. Y., Endroyono, & Yuniarno, E. M. (2019). Combining SentiStrength and Multilayer Perceptron in Twitter Sentiment Classification. Proceedings - 2019 International Seminar on Intelligent Technology and Its Application, ISITIA 2019, 381–386. https://doi.org/10.1109/ISITIA.2019.8937134

Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., & Ettaouil, M. (2016). Multilayer Perceptron: Architecture Optimization and Training. International Journal of Interactive Multimedia and Artificial Intelligence, 4(1), 26. https://doi.org/10.9781/IJIMAI.2016.415

Raza, M. O., Memon, M., Bhatti, S., & Bux, R. (2020). Detecting Cyberbullying in Social Commentary Using Supervised Machine Learning. Advances in Intelligent Systems and Computing, 1130 AISC, 621–630. https://doi.org/10.1007/978-3-030-39442-4_45

Rendalkar, S., & Chandankhede, C. (2018). Sarcasm Detection of Online Comments Using Emotion Detection. Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018, 1244–1249. https://doi.org/10.1109/ICIRCA.2018.8597368

Shah, D. K., Sanghvi, M. A., Mehta, R. P., Shah, P. S., & Singh, A. (2021). Multilabel Toxic Comment Classification Using Supervised Machine Learning Algorithms. Lecture Notes in Networks and Systems, 141, 23–32. https://doi.org/10.1007/978-981-15-7106-0_3

Shtovba, S., Shtovba, O., & Petrychko, M. (2019). Detection of Social Network Toxic Comments with Usage of Syntactic Dependencies in the Sentences. Undefined.