ABUAD INTERNATIONAL JOURNAL OF NATURAL AND APPLIED SCIENCES
ISSN: 2955-1021
AIJNAS Volume 3, Issue 2, 2023 pp 46-51
https://doi.org/10.53982/aijnas.2023.0302.07-j
Copyright ©2023
https://journals.abuad.edu.ng/index.php/aijnas

An Ensemble Approach to Cyberbulling Detection and Prevention on Social Media

Stephen Eyitayo Obamiyi[1], Bukola Badeji-Ajisafe[2], Abiodun Oguntimilehin[3], Treasure Oluwatoyin Adefehinti[4], Oluwatoyin Bunmi Abiola, Toyin Okebule
[1,2,3,5,6*] *Department of Mathematical and Physical Sciences, Afe Babalola University, Ado-Ekiti, Nigeria*
[4*] *Department of Computer Science, Bamidele Olumilua University of Science and Technology, Ikere-Ekiti*
Coresponding Author: Stephen Eyitayo Obamiyi
Email: obamiyise@abuad.edu.ng

Abstract
Over the past decade, digital communication has reached a massive scale globally. Unfortunately, cyberbullying has also seen a significant increase which commensurate with the growth of digital technology, and perpetrators hiding behind the cloak of relative internet anonymity. Studies have shown that cyberbullying leaves a lasting psychological scar on its victims and often have devastating outcome. This has necessitated the development of measures to curb cyberbullying. This study presents one of such measure in the form of an ensemble model for cyberbullying detection. The proposed model features a majority voting ensemble approach to cyberbullying detection using three (3) supervised machine learning classifiers: SVM, NB and K-NN, as base learners. The malignant comment dataset, sourced from Kaggle.com. was used for model building at a split ratio of 70: 30 to achieve maximum model training and evaluation respectively. Evaluation result was based on standard metrics. The proposed ensemble model performed best of all the models implemented, with an accuracy of 95%. It was also observed to be the most consistent classifier across all the metrics considered. This showcased the efficacy of the ensemble model in cyberbullying comments detection.

Keyword: *Majority Voting, Cyberbullying, Cyberbullying detection, Support Vector Machine, Naïve Bayes, K-Nearest Neighbor.*

INTRODUCTION

Recently, people all over the globe make use of various online forums, blogs and social networking sites as a basis for sharing, networking and transfer of knowledge. Teens and youths alike who are usually at the forefront of embracing new technologies have been the most hit with the adverse effects that accompanies these new platforms, one of which is Cyberbullying (Ademiluyi *et al.,* 2022). UNICEF described cyberbullying as the use of digital technology to demean others (Roy & Mali, 2022). The US National Crime Prevention Council (NCPC) described it as the use of the internet, phones and other devices to send texts or post multimedia messages that are intended to hurt or embarrass another person (Williamson, 2010). The phenomenon has been designated a public health threat by the US centre for disease control and prevention (Ferrara *et al.,* 2018), as studies listed its devastating effects on victims to include: depression, suicidal thoughts, anxiety, self-harm and low self esteem (Fisher *et al.,* 2012). In spite of the devastating effects of cyberbullying, incidences of cyberbullying continue to surge in tandem with the growth of online platforms where they are being perpetrated (Gohal *et al.,* 2023). The distributed and anonymous nature of the internet further strengthens the use of these platforms for activities considered highly unethical. The need to keep and protect the mental state and well-being of our teeming youth from of actions of cyberbullies necessitates the development of measures to checkmate the prevalence of cyberbullying and its far-reaching effects. A common measure adopted towards curbing the prevalence of cyberbullying is the development of a system to detect actions and posts that constitute cyberbullying in the cyber space using computer-aided diagnostic approach based on machine learning models.

Machine learning (ML) is a strong AI tool for the development of intelligent systems (Sarker, 2022). It is widely used for modelling diagnostic systems with wide applications in health, IT security, and natural language processing (Shinde & Shah, 2018). ML models learn patterns from historic data, and use the knowledge gained to classify new data with similar features (Taye, 2023). The several models which constitute ML learn and classify data points via different techniques – making their performance vary based on the type of data being classified (Yahyaoui & Yumuşak, 2018). An alternative approach allows for the combination of several models – leveraging their individual strength and weaknesses, to achieve a more stable and accurate result. This approach is known as ensemble learning (Dong *et al.,* 2020). This study seeks to develop a cyberbullying detection system based on ensemble learning.

REVIEW OF RELATED LITERATURE

This section describes previous studies and the associated method they proposed to checkmate the prevalence of cyberbullying in the cyberspace.

Huang *et al.* (2014) presented a study that identify cyberbullying on social media. The proposed approach ranked social media textual features using information gain and proceed to classify the ranked features using versatile classifiers such as; NB, J48, and Bagging.

Al-garadi *et al.* (2016) experimented on cyberbullying identification using diverse ML classifiers such as RF, Naïve Bayes (NB), and SVM based on various extracted features from Twitter such as (tweet content, activity, network, and user).

Perelló et al. (2019) proposed a hybrid model that applies n-gram models and SVM on Twitter messages to determine if they are malignant. The proposed hybrid approach adopted SVM as the classifier for hate speech detection by means of the n-gram model deployed for the extraction of textual features. The stages in the model were to first detect malignant tweets specific to immigrants and women, next was to determine if a malignant tweet is targeted at an individual or a group, and finally to classify malignant tweets as Aggressive or not aggressive for both English and Spanish.

Abro *et al.* (2020) developed a machine learning model to detect cyberbullying via text. The study used the CrowdFlower dataset which contains tweets sourced from Twitter. The tweets which were in textual form was pre-processed by converting uppercase into lowercase and removing URLs, usernames, hashtags and stop-words. Tokenization and lemmatization were also applied to the text. Six classification techniques: NB, SVM, KNN, DT, RF and LR were applied. N-gram with Term Frequency Inverse Document Frequency of records (TF-IDF), Word2vec and Doc2vec feature techniques were also applied. SVM with a combination of bigram and TF-IDF technique showed the best results.

Florio *et al.* (2020) presented a Bidirectional Encoder Representations from Transformers (BERT) pre-trained on Italian hate speech Twitter data, called the AlBERTo, for the prediction of hate speech. The AlBERTo is made up of an encoder layer that performs sentence-level feature representation and a decoder layer that generates a binary output denoting hate or non-hate speech. Experimental result found the AlBERTo model better in comparison with linear SVM when evaluated with standard metrics.

Faris *et al.* (2020) presented a study that address the hate speech detection problem in Arabic language using a combination of word embedding and deep learning approach. The proposed approach achieved word embedding on hate speech dataset sourced from Twitter using LSTM. This helps to get tweet features on diverse topics at the Arabic region. The features extracted was then classified using CNN. On experimentation, the proposed LSTM-CNN model was observed to have a superior classification accuracy categorizing tweets as either hate or normal when compared to other models.

Ayo *et al.* (2020) in their study presented a hybrid embedding model which is enhanced with a topic inference method, and an improved cuckoo search neural network for detecting hate speech in Twitter data. The presented approach implements a hybrid embedding technique that includes Term Frequency-Inverse Document Frequency (TF-IDF) for word-level feature extraction and LSTM for sentence-level feature extraction. The extracted features from the hybrid embeddings then fed into an improved cuckoo search neural network for hate speech detection. The study categories tweets into one of the following categories: hate speech, offensive language or neither.

Muneer & Fati (2020) implemented several unique classifiers, namely AdaBoost (ADB), Light Gradient Boosting Machine (LGBM), SVM, RF, Stochastic Gradient Descent (SGD), Logistic Regression (LR), and MNB, for the detection of tweets that constitute cyberbullying. This study extracted features using Word2Vec and TF-IDF methods

Dalvi et al. (2021) used SVM and Random Forests (RF) models with TF-IDF for feature extraction for detecting cyberbullying in tweets. Although SVM in these models achieved high performance, the model complexity increases when the class labels are increased.

Balakrishnan *et al.* (2020) utilized different ML algorithms such as RF, NB, and J48 to detect cyberbullying events from tweets and classify tweets to different cyberbullying classes such as aggressors, spammer, bully, and normal. The study concluded that the emotional feature does not impact the detection rate. Despite its efficiency, this model is limited to a small dataset with fewer class labels.

Murshed *et al.* (2022) proposed DEA-RNN, a hybrid deep learning model to detect cyberbullying on Twitter. The proposed DEA-RNN model combines Elman type Recurrent Neural Networks (RNN) with an optimized Dolphin Echolocation Algorithm (DEA) for tuning the parameters of Elman RNN, and reducing training time. The model was evaluated using a dataset of 10000 tweets, and its performance compared with Bi-directional long short-term memory (Bi-LSTM), RNN, SVM, Multinomial Naive Bayes (MNB), and Random Forests (RF). The experimental results show that the DEA-RNN outperformed the considered existing approaches.

METHODOLOGY

To achieve an ensemble model for cyberbullying detection in the cyberspace, the study followed some fundamental steps ranging from extensive survey of related literature, to choosing an appropriate ensemble for the detection model. Figure 1 describes the framework for the proposed model.

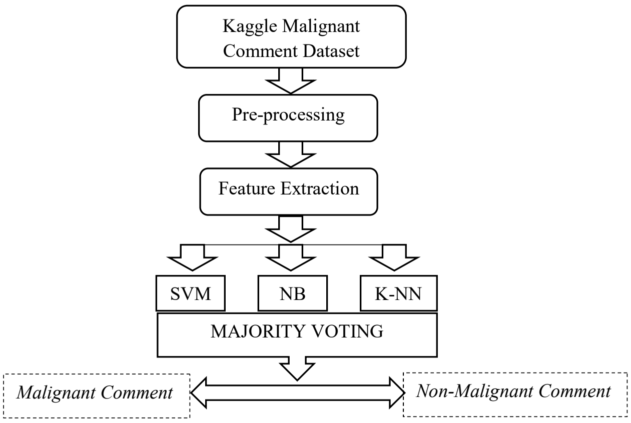

Figure 1: Proposed cyberbullying detection model framework.

Dataset Acquisition and Description

Malignant comment classification dataset was sourced from Kaggle.com and used as the design set for this study. The dataset is composed of approximately 150,000 records. These constituted the design set for the study. All records in the dataset in made up of 8 attributes which includes; 'Id', 'Comments', 'Malignant', 'Highly malignant', 'Rude', 'Threat', 'Abuse' and 'Loathe'. Description of these fields is presented in table 1. Each attribute other than the Id and comment holds a value of either 0 or 1, denoting NO and YES to the attribute.

Table 1: Description of dataset attributes

S/N	ATTRIBUTE	TYPE	DESCRIPTION
1	ID	Ordinal	Unique Ids associated with each comment text given
2	Comment text	Nominal	Comments extracted from various social media platforms
3	Loathe	Ordinal	Comments which are hateful and loathing in nature
4	Abuse	Ordinal	Comments that are abusive in nature
5	Threat	Ordinal	Comments that that constitute threat to someone
6	Rude	Ordinal	Comments that are very rude and offensive
7	Highly Malignant	Ordinal	Comments that are highly malignant and hurtful
8	Malignant	Ordinal	Comments that are malignant and hurtful

Dataset Preprocessing

Data preprocessing helps to make the dataset suitable for the development of the cyberbullying detection model. The preprocessing measures taken in the study include:

Decapitalization: this involve rendering all textual features to its equivalent lowercase form

Removal of stop words, punctuations and names: stop words which are commonly used words and deemed unimportant to the detection model's performance were removed to help the model concentrate on more important words. Examples of such stop words include: "a", "on", and etc. Similarly, all punctuation marks and names of persons mentioned in the textual comment were removed.

Feature Extraction using Bag of Words (BoWs)

The bag of words model provides a way of representing textual data in numbers when modeling text with machine learning algorithms. It extracts feature sets from text during data preprocessing. The approach involves breaking a textual data down into a list of disparate words and noting the frequency of each word as used in the data.

Machine Learning Classifier for Cyberbullying Detection.

The study has evaluated the performance of three machine learning classifiers: Support Vector Machine, Naïve Bayes and K nearest neighbor in detecting cyberbullying from the malignant comment dataset sourced from Kaggle.com. the study went further to ensemble the three classifiers using majority voting to boost detection rate. The models are explained thus.

Support Vector Machine

Support Vector Machine (SVM) is a versatile supervised machine learning algorithm widely used for classification and regression problem. SVM is beloved in the research community for its ability to perform significantly with less computational power. SVM plot its data item as a point in an n-dimensional space. It uses the value of each feature to map the feature to a specific coordinate, then, attempt to achieve classification by obtaining an optimum hyper-plane that best separates the features to their individual class (Durgesh & Lekha, 2010).

Naïve Bayes

The Naïve Bayes classifier is a supervised machine learning algorithm that is built upon the Bayes probability theory, and widely used for classification tasks, like text classification (Joyce & Deng, 2019). It is a generative learning algorithm that seeks to model the distribution of inputs of a given class or category. Naïve Bayes is so beloved as it requires a small amount to training data to estimate the necessary parameters. They are also extremely fast when compared to more sophisticated classification models. The Bayes conditional probability upon which Naïve Bayes algorithm is built is described in equ 1:

$$P(C/x) = \frac{P(x/C).\ P(C)}{P(x)} \qquad \text{equ. 1}$$

Where: $P(C/x)$ is the posteriori probability of target class C, given attribute x
$P(C)$is the priori probability of target class C
$P(x/C)$ is the probability of attribute x, given class C
$P(x)$is the priori probability of attribute x

K-Nearest Neighbor

The k-nearest neighbors otherwise known as k-NN, is a non-parametric, supervised learning algorithm that classifies on the basis of the proximity of the individual data points to be classified. K-NN is widely used for both regression and classification problems. It performs learning and prediction analysis of a given problem based on a distance function (usually the Euclidean distance) and a voting function, with the number of votes limited to K (Altay & Ulas, 2018). Prediction in K-NN is purely based on neighbor data values without any assumption on the dataset. The Euclidean distance between two sample can be computed using:

$$Dist(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \text{equ. 2}$$

Where: $Dist(x, y)$: Euclidean distance between vector x and y
x_i: testing data I, with I = 1, 2, … , n
y_i: training data I, with I = 1, 2, … , n
n: amount of attributes

Majority Voting Ensemble

Majority voting is a machine learning technique that aggregates the predictions of multiple other models to make a more accurate prediction based on the plurality of the outcome of the classifiers combined. It is both a homogeneous and heterogeneous ensemble learning technique that enables the base classifiers to each contribute a single vote to the final outcome of the model. The class with the highest votes from the base classifiers is returned as the correct class for the data instance being classified. Figure 2 shows the majority voting model architecture. Also, given three base classifiers () with predictions . Then, the final prediction can be obtained as:

$$P_F = Mode(P_1, P_2, P_3) \qquad \text{equ. 3}$$

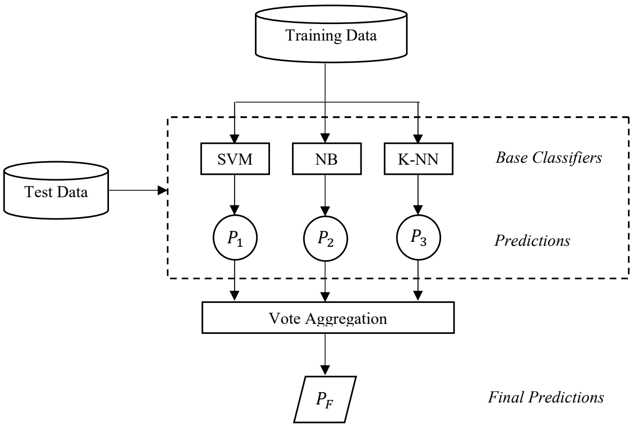

Figure 2: Architecture of the majority voting classifier

Model Evaluation Metrics

The various models implemented were evaluated using standard metrics. A binary classifier labels all data elements in a test dataset with a 0 or 1. Classification result is falls into one of the following categories: True positive (TP), True negative (TN), false positive (FP), and false negative (FN). The following equations were used to computer the classifier's accuracy, precision, recall and F-Score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} \qquad \text{equ. 4}$$

$$Precision = \frac{TP}{TP+FP} \qquad \text{equ. 5}$$

$$Recall = \frac{TP}{TP+FN} \qquad \text{equ. 6}$$

$$F1\ Score = 2 * \left(\frac{Precision*Recall}{Precision+Rec}\right) \qquad \text{equ. 7}$$

RESULT AND DISCUSSION

A comprehensive summary of the performance of the various classifiers is presented in table 2 and table 3. Table 2 gives a summary of the correct classifications and misclassifications by the individual classifiers while table 3 presents the evaluation result for the classifiers based on the four metrics earlier mention.

Table 2: Confusion matrix table for the classifiers.

Confusion Matrix Table **0** **1**			**Predicted**	
Actual	SVM	0	54871	2770
		1	1264	4924
	NB	0	55887	1754
		1	1623	4565
	K-NN	0	57388	253
		1	4402	1786

Table 3: Evaluation Result of the Classifiers

Classifier	Accuracy	Precision	Recall	F1-Score
SVM	0.94	0.64	0.80	0.71
NB	0.95	0.72	0.74	0.73
K-NN	0.93	0.88	0.29	0.43
Voting	0.96	0.81	0.70	0.75

Table 3 expressly showed that the voting ensemble has the overall best accuracy with a score of 0.96. Naïve bayes algorithm is next with an accuracy score of 0.95. SVM has an accuracy score of 0.94 and K-NN has the least accuracy, with an accuracy score of 0.93. In terms of Precision, K-NN was observed to have the best precision, with a score of 0.88, followed by the voting ensemble which has a precision of 0.81, Naïve Bayes with a precision of 0.72 and SVM with a Precision of 0.64. Across the various metrics used for evaluation, the classifiers performed in a staggering fashion as observed in table 3. The voting ensemble though was observed to have a more stable performance across all the metrics used for the model development.

Figure 3 Presents a graphical representation of the performance of all classifiers used in the study for easy comparison.

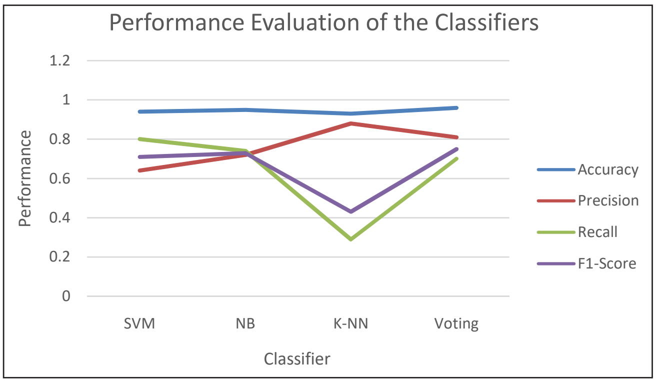

Figure 3: Graphical presentation of the model performance

CONCLUSION

Cyberbullying is a public menace that significantly contribute to the deterioration its victim's mental health, resulting to low self esteem amongst it several other effects. This study has presented a majority voting ensemble approach to cyberbullying detection using three (3) supervised machine learning classifiers: SVM, NB and K-NN, as base learners. The proposed model was trained and evaluated on the malignant comment dataset which was sourced on Kaggle.com. The dataset which contains 150,000 instances of the data was split at a ratio of 70: 30 to achieve maximum model training and evaluation respectively. Evaluation result was based on standard metrics and showed the efficacy of the ensemble model in cyberbullying comments detection, as it was the most consistent classifier across all the metrics considered.

Even though the ensemble model could accurately classify up to 96%, the precision, at 81% still needs to perform more. Further studies could look into the effects of each data attribute on machine learning model performance and come up with ways of optimizing textual data preprocessing to ensure minimal loss of vital information. Superior classification models could also be implemented to boost cyberbullying detection rate.

REFERENCES

Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, *11*(8).

Ademiluyi, A., Li, C., & Park, A. (2022). Implications and preventions of cyberbullying and social exclusion in social media: systematic review. *JMIR formative research*, *6*(1), e30286.

Al-Garadi, M. A. Varathan, K. D. & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network,'' Comput. Hum. Behav., vol. 63, pp. 433–443.

Altay, O., & Ulas, M. (2018, March). Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. In *2018 6th international symposium on digital forensic and security (ISDFS)* (pp. 1-4). IEEE.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks. *International Journal of Intelligent Computing and Cybernetics*, *13*(4), 485-525.

Balakrishnan, V., Khan, S. & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning,'' Comput. Secur., vol. 90, Art. no. 101710,

Dalvi, R. R., Chavan, S. B. & Halbe, A. (2021). Detecting a Twitter cyberbullying using machine learning,'' Ann. Romanian Soc. Cell Biol., vol. 25, no. 4, pp. 16307–16315.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241-258.

Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. *Journal of theoretical and applied information technology*, *12*(1), 1-7.

Faris, H., Habib, I.A.M. and Castillo, P.A. (2020), "Hate speech detection using word embedding and deep learning in the Arabic language context", *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020),* pp. 453-460.

Ferrara, P., Ianniello, F., Villani, A., & Corsello, G. (2018). Cyberbullying a modern form of bullying: let's talk about this health and social problem. *Italian journal of pediatrics*, *44*(1), 1-3.

Fisher, H. L., Moffitt, T. E., Houts, R. M., Belsky, D. W., Arseneault, L., & Caspi, A. (2012). Bullying victimisation and risk of self harm in early adolescence: longitudinal cohort study. *Bmj*, *344*, e2683.

Florio, K., Basile, V., Polignano, M., Basile, P. and Patti, V. (2020), "Time of your hate: the challenge of time in hate speech detection on social media", *Applied Sciences*, Vol. 10 No. 12, p. 4180.

Gohal, G., Alqassim, A., Eltyeb, E., Rayyani, A., Hakami, B., Al Faqih, A., ... & Mahfouz, M. (2023). Prevalence and related risks of cyberbullying and its effects on adolescent. *BMC psychiatry*, *23*(1), 39.

Huang, Q., Singh, V. K. & Atrey, P. K. (2014). Cyber bullying detection using social and textual analysis," *in Proc. 3rd Int. Workshop Socially-Aware Multimedia (SAM),* pp. 3–6.

Joyce, B., & Deng, J. (2019). Sentiment Analysis Using Naive Bayes Approach with Weighted Reviews-A Case Study. In *2019 IEEE Global Communications Conference (GLOBECOM)* (pp. 1-6). IEEE.

Muneer, A. and Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. *Futur. Internet*, 12(11), pp. 1–21, 2020.

Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access*, *10*, 25857-25871.

Perelló, C., Tomás, D., Garcia-Garcia, A., Garcia-Rodriguez, J., & Camacho-Collados, J. (2019). UA at SemEval-2019 task 5: setting a strong linear baseline for hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 508-513).

Roy, P. K., & Mali, F. U. (2022). Cyberbullying detection using deep transfer learning. *Complex & Intelligent Systems*, *8*(6), 5449-5467.

Sarker, I. H. (2022). Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, *3*(2), 158.

Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1-6). IEEE.

Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers*, *12*(5), 91.

Williamson, R. (2010). Cyberbullying. *Education Partnerships, Inc, pp 1-10*

Yahyaoui, I., & Yumuşak, N. (2018). Machine learning techniques for data classification. In *Advances in renewable energies and power technologies* (pp. 441-450). Elsevier.