Social Media Cyberbullying Detection using Deep Learning

C.Sahithi , Ch. Siddhu Vinay, Ch.Tharun reddy, C. Ashritha, Dr.S.Satyanarayana

School of Engineering, Malla Reddy University, Secunderabad – 500010, Telangana, India;
chennurusahithi2002@gmail.com,siddhuvinay0511@gmail.com,tharunreddychinthareddy@gmail.com,
ashrithachincholi@gmail.com

Abstract

Cyberbullying is a disturbing online misbehaviour with troubling consequences. It appears in different forms, and in most of the social networks, it is in textual format. Automatic detection of such incidents requires intelligent systems. Most of the existing studies have approached this problem with conventional machine learning models and the majority of the developed models in these studies are adaptable to a single social network at a time. In recent studies, deep learning based models have found their way in the detection of cyberbullying incidents, claiming that they can overcome the limitations of the conventional models, and improve the detection performance.

1. Introduction

Cyberbullying, a prevalent form of harassment in the digital age, has become a concerning issue with the rapid advancement of technology and the widespread use of social media platforms. It involves the use of electronic communication tools to intimidate, threaten, or harm individuals, often with the intention to cause emotional distress or damage their reputation. To combat this pervasive problem, researchers and experts have turned to deep learning techniques as a means to detect and prevent cyberbullying.Deep learning, a subset of machine learning, employs artificial neural networks to process large amounts of data and extract meaningful patterns. In the context of cyberbullying, deep learning models can be trained using vast datasets containing instances of cyberbullying and non-cyberbullying interactions. These models can then learn to recognize various forms of online harassment, such as explicit language, derogatory remarks, threats, or targeted attacks.By leveraging deep learning algorithms, these models can analyze text, images, and videos posted on social media platforms or other online channels. Natural Language Processing (NLP) techniques enable the identification of offensive or abusive language, while computer vision algorithms allow the detection of harmful images or videos. The models can also consider contextual factors like the relationship between the involved parties and the historical behavior of the user to provide a more comprehensive analysis.Through continuous training and refinement, deep learning models can improve their accuracy in identifying cyberbullying incidents. These models can be integrated into social media platforms, messaging apps, or other online spaces to automatically flag potentially harmful content, alert users, and prompt intervention from human moderators. The goal is to create a safer online environment by promptly identifying and addressing instances of cyberbullying, thereby reducing the potential harm inflicted on individuals.

2. Related work

There are many approaches that proposes systems which can detect cyberbullying automatically with high accuracy. First one is author Nandhini et al. [3] have proposed a model that uses Naïve Bayes machine learning approach and by their work they achieved 91% accuracy and got their dataset from MySpace.com, and then they proposed another model [4] Naïve Bayes classifier and genetic operations (FuzGen) and they achieved 87% accuracy. Another approach by Romsaiyud et al. [5] they enhanced the Naïve Bayes classifier for extracting the words and examining loaded pattern clustering and by this approach they achieved 95.79% accuracy on datasets from Slashdot, Kongregate, and MySpace. However, they have a problem that the cluster processes doesn't work in parallel. Moreover, in the approach proposed by Bunchanan et al. [6] they used War of Tanks game chat to get their dataset and manually classified them and then compared them to simple Naïve classification that uses sentiment analysis as a feature, their results were poor when compared to the manually classified results. Furthermore, Isa et al. [7] proposed an approach after getting their dataset from kaggle they used two classifier Naïve Bayes and SVM. The Naïve Bayes classifier yielded average accuracy of 92.81% while SVM with poly kernel yielded accuracy of 97.11%, but they did not mention their training or testing size of the dataset, so the results may not be credible. Another Approach by Dinakar et al. [8] that aimed to detect explicit bullying language pertaining to (1) Sexuality, (2) Race & Culture and (3) intelligence, they acquired their dataset from YouTube comment section. After applying SVM and Naïve Bayes classifiers, SVM yielded accuracy of 66% and Naïve Bayes 63%. Moving on to Di Capua et al. [9], they proposed a new way for cyberbullying detection by adopting an unsupervised approach, they used the classifiers inconsistently over their dataset, applying SVM on FormSpring and achieving 67% on recall, applying GHSOM on YouTube and achieving

60% precision, 69% accuracy and 94% recall, applying Naïve Bayes on Twitter and achieving 67% accuracy.

Additionally, Haidar et al. [10] proposed a model to detect cyberbullying but using Arabic language they used Naïve Bayes and achieved 90.85% precision and SVM achieved 94.1% as precision but they have high rate of false positive also the are work on Arabic language.

Another type of approaches using Deep Learning and Neural Networks. One of the proposed methods is Zhang et al. [11] in their paper uses novel pronunciation based convolution neural network (PCNN), thereby alleviating the problem of noise and bullying data sparsity to overcome class imbalance. 1313 messages from twitter, 13,000 messages from formspring.me. Accuracy of the twitter dataset wasn't calculated due to it being imbalanced. While Achieving 56% on precision, 78% recall and 96% accuracy, while achieving high accuracy their dataset was unbalanced, so that gives false results and that reflects in precision score which is 56%. The authors Nobata et al. [12] showed that using abusive language has increased recently, They used a framework called Vowpal wabbit for classification, and they also developed a supervised classification methodology with NLP features that outperform deep learning approach, The F-Score reached 0.817 using dataset collected from comments posted on Yahoo News and Finance.Zhao et al.

3. Proposed Methodology

Cyberbullying detection using MLP (Multi-Layer [illegible] network, can effectively learn patterns and relationships in textual data, enabling accurate classification of cyberbullying instances. Here's how MLP can be utilized for cyberbullying detection:

- Tokenization: In this part we take the text as sentences or whole paragraphs and then output the entered text as separated words in a list.

Lowering text: This takes the list of words that got out of the tokenization and then lower all the letters Like: 'THIS IS AWESOME' is going to be 'this is awesome'.

Stop words and encoding cleaning: This is an essential part of the preprocessing where we clean the text from those stop words and encoding characters like \n or \t which do not provide a meaningful information to the classifiers.

Word Correction: In this part we used Microsoft Bing word correction API [24] that takes a word and then return a JSON object with the most similar words and the distance between these words and the original word.

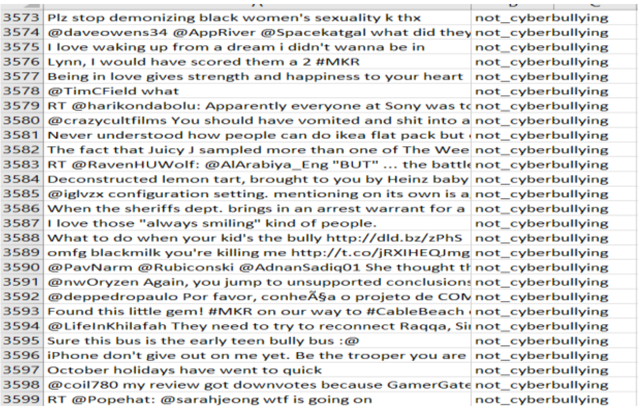
Dataset Preparation:

We Collect a well-labeled dataset that includes instances of cyberbullying and non-cyberbullying interactions. Ensure that the dataset is diverse, covering various forms of cyberbullying behaviors and contexts.

Data set consist of two columns and 476953 rows with contains various text and classification such as cyberbullying and nnon—cyberbullying .

1.Raw Data:

The raw data is basically the unprocessed or we can say the data which is not completely ready to be used for any information extraction. It is also known as the source data which has not been gone through any processing technique whether manually or through any algorithm or any automated machine. The below mentioned is the primary data set which has been taken from kaggle data sets. The dataset which is available has following tuples (476953) with attributes (2). The data set given in Figure 2 is in the format of seconds, so the next step is to process the data and to do this the certain algorithms and processes are used to make our data relevant to use.



3573	Plz stop demonizing black women's sexuality k thx	not_cyberbullying
3574	@daveowens34 @AppRiver @Spacekatgal what did they	not_cyberbullying
3575	I love waking up from a dream i didn't wanna be in	not_cyberbullying
3576	Lynn, I would have scored them a 2 #MKR	not_cyberbullying
3577	Being in love gives strength and happiness to your heart	not_cyberbullying
3578	@TimCField what	not_cyberbullying
3579	RT @harikondabolu: Apparently everyone at Sony was to	not_cyberbullying
3580	@crazycultfilms You should have vomited and shit into a	not_cyberbullying
3581	Never understood how people can do ikea flat pack but	not_cyberbullying
3582	The fact that Juicy J sampled more than one of The Wee	not_cyberbullying
3583	RT @RavenHUWolf: @AlArabiya_Eng "BUT" ... the battle	not_cyberbullying
3584	Deconstructed lemon tart, brought to you by Heinz baby	not_cyberbullying
3585	@iglvzx configuration setting. mentioning on its own is a	not_cyberbullying
3586	When the sheriffs dept. brings in an arrest warrant for a	not_cyberbullying
3587	I love those "always smiling" kind of people.	not_cyberbullying
3588	What to do when your kid's the bully http://dld.bz/zPhS	not_cyberbullying
3589	omfg blackmilk you're killing me http://t.co/jRXIHEQJmg	not_cyberbullying
3590	@PavNarm @Rubiconski @AdnanSadiq01 She thought th	not_cyberbullying
3591	@nwOryzen Again, you jump to unsupported conclusions	not_cyberbullying
3592	@deppedropaulo Por favor, conheÃ§a o projeto de COM	not_cyberbullying
3593	Found this little gem! #MKR on our way to #CableBeach	not_cyberbullying
3594	@LifeInKhilafah They need to try to reconnect Raqqa, Si	not_cyberbullying
3595	Sure this bus is the early teen bully bus :@	not_cyberbullying
3596	iPhone don't give out on me yet. Be the trooper you are	not_cyberbullying
3597	October holidays have went to quick	not_cyberbullying
3598	@coil780 my review got downvotes because GamerGate	not_cyberbullying
3599	RT @Popehat: @sarahjeong wtf is going on	not_cyberbullying

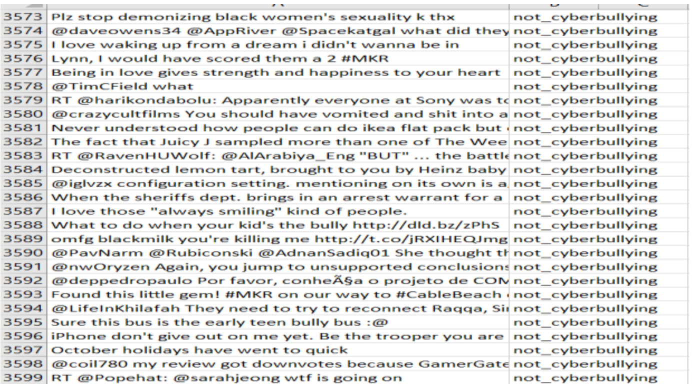
Figure 2. Dataset
https://www.kaggle.com/datasets/madhubalaji/cyber-bullying-tweets

3.Data Predictions:

In this phase predictions are made out by the implementation of proper algorithms. In this paper the two algorithms are used for the data prediction i.e.MLP and other is SVM (support vector machine). Both algorithms depict different predictions. The results obtained are in the form of RMSE and MAPE values. These predictions are further utilized for data visualization.

2.Preprocessing:

In the phase of preprocessing the main objective is to select the required data which include following attributes text like age , enthnicity, gender,religion .The preprocessing here is carried out through visualization technique and training is done through the recurrent neural network. It is a must step in every proposed methodologies for the development of unique and more accurate model. Data processing is basically the data filtering so as to make it more convenient to use. In these proposed methodologies after the dataset is trained with the inclusion of following attributes is depicted below. The attributes given in Figure 3 are being used for the prediction.



3573	Plz stop demonizing black women's sexuality k thx	not_cyberbullying
3574	@daveowens34 @AppRiver @Spacekatgal what did they	not_cyberbullying
3575	I love waking up from a dream i didn't wanna be in	not_cyberbullying
3576	Lynn, I would have scored them a 2 #MKR	not_cyberbullying
3577	Being in love gives strength and happiness to your heart	not_cyberbullying
3578	@TimCField what	not_cyberbullying
3579	RT @harikondabolu: Apparently everyone at Sony was t	not_cyberbullying
3580	@crazycultfilms You should have vomited and shit into a	not_cyberbullying
3581	Never understood how people can do ikea flat pack but	not_cyberbullying
3582	The fact that Juicy J sampled more than one of The Wee	not_cyberbullying
3583	RT @RavenHUWolf: @AlArabiya_Eng "BUT" ... the battl	not_cyberbullying
3584	Deconstructed lemon tart, brought to you by Heinz baby	not_cyberbullying
3585	@iglvzx configuration setting. mentioning on its own is a	not_cyberbullying
3586	When the sheriffs dept. brings in an arrest warrant for a	not_cyberbullying
3587	I love those "always smiling" kind of people.	not_cyberbullying
3588	What to do when your kid's the bully http://dld.bz/zPhS	not_cyberbullying
3589	omfg blackmilk you're killing me http://t.co/jRXIHEQJmg	not_cyberbullying
3590	@PavNarm @Rubiconski @AdnanSadiq01 She thought t	not_cyberbullying
3591	@nwOryzen Again, you jump to unsupported conclusion	not_cyberbullying
3592	@deppedropaulo Por favor, conheÃ§a o projeto de CO	not_cyberbullying
3593	Found this little gem! #MKR on our way to #CableBeach	not_cyberbullying
3594	@LifeInKhilafah They need to try to reconnect Raqqa, Si	not_cyberbullying
3595	Sure this bus is the early teen bully bus :@	not_cyberbullying
3596	iPhone don't give out on me yet. Be the trooper you are	not_cyberbullying
3597	October holidays have went to quick	not_cyberbullying
3598	@coil780 my review got downvotes because GamerGate	not_cyberbullying
3599	RT @Popehat: @sarahjeong wtf is going on	not_cyberbullying

Processed dataset

In this type of methodology basically the prediction is based on the neural network techniques, result are obtained by the neural network mechanism. In the neural networks the connection between the nodes are in the form of directed cycles which are the recurrent neural network. The algorithm which is used for training purpose is Back propagation algorithm in which the parameters are shared in all the steps for better and efficient result.

4.Visualization:

Visualization is the process of analyzing of obtained result, in this phase the predicted wind data is visualized and the result obtained in the form of RMSE and MAPE properly analyzed to generate the predicted graphs, bar graphs and other necessary pictorial representations.

3.1 MLP based Methodology

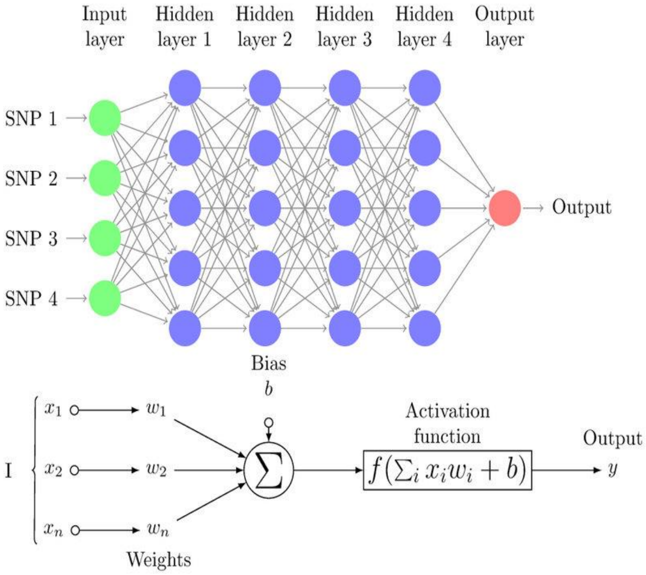
In the proposed methodology the first task is to gather the available data for implementation, collection of this type of data is based on various parameters. Here parameters are the basic environmental parameters. In pre-processing phase second task is to reduce the dimensionality of the data through data visualisation technique. The main objective of data visualisation is to select the necessary parameters from the complete dataset. The new set of variables obtained are now trained using recurrent neural network algorithm (back propagation). After the training of data is completed the third task is to apply MLP over the trained data to obtain the predictions and to generate the error rate (RMSE value).

A multilayer perceptron (MLP) is a type of artificial neural network commonly used for supervised learning tasks such as classification. MLPs are composed of multiple layers of interconnected neurons, where each neuron receives input from multiple neurons in the previous layer and produces an output that is passed to multiple neurons in the next layer.

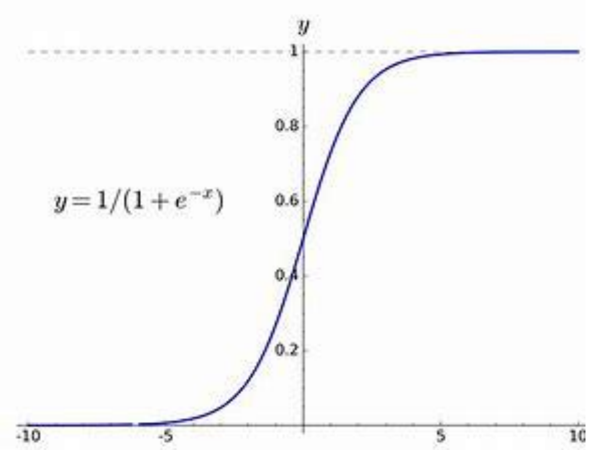
Multi-layer perception is also known as MLP. It is fully connected dense layers, which transform any input dimension to the desired dimension. A multi-layer perception is a neural network that has multiple layers. To create a neural network we combine neurons together so that the outputs of some neurons are inputs of other neurons.

Multi-layer perception is also known as MLP. It is fully connected dense layers, which transform any input dimension to the desired dimension. A multi-layer perception is a neural network that has multiple layers. To create a neural network we combine neurons together so that the outputs of some neurons are inputs of other neurons.

In the multi-layer perceptron diagram above, we can see that there are three inputs and thus three input nodes and the hidden layer has three nodes. The output layer gives two outputs, therefore there are two output nodes. The nodes in the input layer take input and forward it for further process, in the diagram above the nodes in the input layer forwards their output to each of the three nodes in the hidden layer, and in the same way, the hidden layer processes the information and passes it to the output layer.



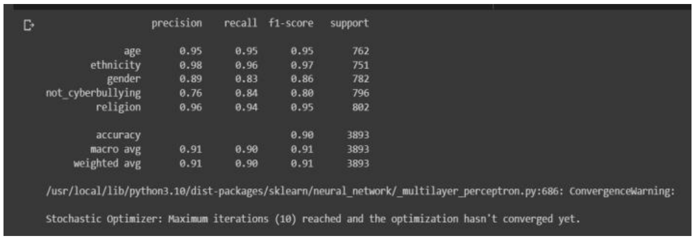
Working of MLP



SIGMOID FUNCTION

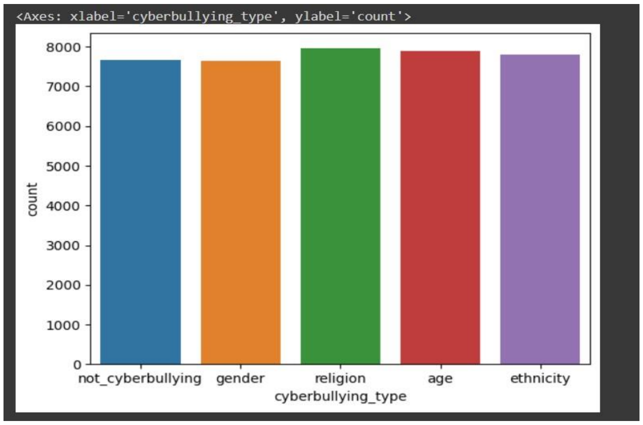
4.Results

MLP is applied on the data set from kaggle datasets which includes attributes in terms of age , ethnicity , gender , religion and Not cyberbullying The predicted result is quite different in terms RMSE value. The error rate in MLP is 0.427 while in case of SVM according to existing projects it is 0.768. Figure 8 shows the accuracy of performane matrix obtained while the model implementation.



```
                   precision    recall  f1-score   support

              age       0.95      0.95      0.95       762
        ethnicity       0.98      0.96      0.97       751
           gender       0.89      0.83      0.86       782
not_cyberbullying       0.76      0.84      0.80       796
         religion       0.96      0.94      0.95       802

         accuracy                           0.90      3893
        macro avg       0.91      0.90      0.91      3893
     weighted avg       0.91      0.90      0.91      3893

/usr/local/lib/python3.10/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:686: ConvergenceWarning:

Stochastic Optimizer: Maximum iterations (10) reached and the optimization hasn't converged yet.
```

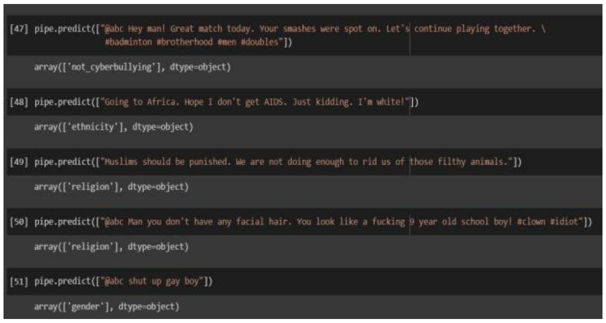
MLP PERFORMANCE



```
<Axes: xlabel='cyberbullying_type', ylabel='count'>
```

COUNT OF TEXT TYPE

Figure given above shows the graph depicting the count of each type of words in training of data done at implementation phase of these methodologies. The followed graph shows the actual count of types of words in case of MLP methodology.At the time of training and testing, certain type of losses occurs, which is due to the uneven occurrence of data. These losses occur due to the glitches in the algorithms.



```
[47] pipe.predict(["@abc Hey man! Great match today. Your smashes were spot on. Let's continue playing together. \
                   #badminton #brotherhood #men #doubles"])

     array(['not_cyberbullying'], dtype=object)

[48] pipe.predict(["Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!"])

     array(['ethnicity'], dtype=object)

[49] pipe.predict(["Muslims should be punished. We are not doing enough to rid us of those filthy animals."])

     array(['religion'], dtype=object)

[50] pipe.predict(["@abc Man you don't have any facial hair. You look like a fucking 9 year old school boy! #clown #idiot"])

     array(['religion'], dtype=object)

[51] pipe.predict(["@abc shut up gay boy"])

     array(['gender'], dtype=object)
```

Fig shows the predictions made by the Random forest algorithm using pipeline .It detects the text iiks cyberbullied or non-cyberbullied text where if it is cyberbullying is will display the type of the bullied text it is such as age based , religion , gender , ethnicity.

Generally, the evaluation of classifiers is done using several evaluation matrices depends on the confusion matrix. Among of those criteria are Accuracy, precision, recall and f-score. They are calculated according to the following equations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$(3) \quad Recall = \frac{TP}{TP+FN}$$

$$F - Score = \frac{2*precision*recall}{precision+recall}$$

5.Conclusion

In this paper, we proposed an approach to detect cyberbullying using deep learning techniques. We evaluated our model using multiple classifiers like SVM,LogisticRegression,Randomforest,Naïve Bayes and Neural Network and we used TFIDF and sentiment analysis algorithms for features extraction. The classifications were evaluated on different n-gram language models. We achieved 80% accuracy using Logistic regression while using both TFIDF and sentiment analysis together. We found that our Neural Network performed better than the SVM classifier as it also achieves average f-score 92% while the SVM achieves average f-score 87%. Furthermore, we compared our work with another related work that used the same dataset, finding that our Neural Network outperformed their classifiers in terms of accuracy and f-score. By achieving this accuracy, our work is definitely going to improve cyberbullying detection to help people to use social media safely. However, detecting cyberbullying pattern is limited by the size of training data. Thus, a larger cyberbullying data is needed to improve the performance. Hence, deep learning techniques will be suitable in the larger data as they are proven to outperform machine learning approaches over larger size data

6.References

[1] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385, 2008.

[2] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE, 2011.

[3] B Nandhini and JI Sheeba. Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, page 20. ACM, 2015.

[4] B Sri Nandhini and JI Sheeba. Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45:485–492, 2015.

[5] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. Automated cyberbullying detection using clustering appearance patterns. In *Knowledge and*

Smart Technology (KST), 2017 9th International Conference on, pages 242– 247. IEEE, 2017.

[6] Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security*, 76:197–213, 2018.

[7] Sani Muhamad Isa, Livia Ashianti, et al. Cyberbullying classification using text mining. In *Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on*, pages 241–246. IEEE, 2017.

[8] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

[9] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. Unsupervised cyber bullying detection in social networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 432–437. IEEE, 2016.

[10] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. *Advances in Science, Technology and Engineering Systems Journal*, 2(6):275–284, 2017.