Using Machine Learning, Image Processing & Neural Networks to Sense Bullying in K-12 Schools

Lalit Kumar Palash Goyal Rishav Kumar

lalit.k1997@gmail.com palashgoyal1608@mountcarmeldelhi.com karanmalik2000@gmail.com

Abstract

We all have heard about bullying and we know that it is an immense challenge that schools have to tackle. Many lives have been ruined due to bullying and the fear it implants into students' mind has caused many of them to go into depression which can lead to suicide. Traditional methods [1] need to be accompanied with modern technology to make the method more effective and efficient. If real time alerts are to school staff, they can identify the perpetuator and extricate the victim swiftly. It this proposed method an AI based solution is implemented to monitor students using standard school surveillance technologies and CCTV to maintain a decorum and safe environment in the school premise. Also the proposed method utilizes other unstructured sources such as attendance records, social media activity and general nature of the students to deliver quick response. Artificial Intelligence (AI) techniques like Convolutional Neural Networks (CNN), which includes image processing and multivariate linear and logistic regression models for classification, are used. Further, the model included sentiment analysis to identify commonly used abuse terms and noisy labels to improve overall model accuracy. The model has been trained and validated with the realistic data from all the sources mentioned and has achieved the classification accuracy of 87% for detecting any sign of bullying.

***Index Terms*-- AI, CNN, Class Entropy Loss, Data-pre-processing, Data pipeline, Facial Recognition, NLP, NN, RELU, Sigmoid, Sentiment analysis, VGG16.**

Introduction

What is bullying? For starters, bullying is aggressive behaviour among teenagers. This comes from many factors which involve ego, power imbalance, upbringing etc. This behaviour has no bounds and can impact a victim's physical and mental health conditions. Bullying can take many forms from verbal to physical. With the introduction of social media, social bullying, a new type of bullying which involves harming one's reputation or relationships, has sprung up. A study was conducted by Symantec Reports with the help of parents of many victims. They noted that almost 24% of the students were involved in some shape or form of bullying [2]. Despite such a high percentage of victims, most methods to keep bullying in check have not borne any fruit and mitigation against bullying remains an enigma.

The traditional methods with their limited use of technology are proved not be very effective. It is time to integrate modern technology to develop more intelligent solutions. The proposed method in this paper will use specific elements of AI tools which exploit the already available school infrastructure to make it a new means of keeping taps on bullying.

Methodology of research and solution development:

First of all, we have identified all the physical, mental, emotional, social and cyber bullying types and parameters that are prevalent. For this extensive study of existing literature has been done [1] [2] [3]. Subsequently these parameters were studied to identify how they can be objectively analysed. This involved identifying the data sources that can give input signals for that bullying parameter. Once the data sources were identified actual data was recovered from these sources. Again, existing literature was studied [4], [5], [6] extensively to identify AI/ML algorithms that can be used to classify them with binary outputs linked to whether this is a bullying situation or not. The authors then put together all these elements to develop a working model of their solution with a proof of concept done in a school. The research methodology followed in this is as given below in figure 1.

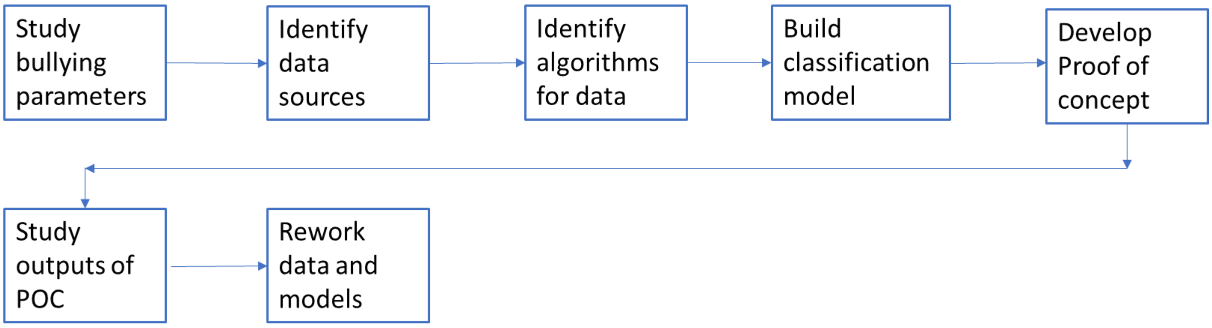

Fig. 1: Research methodology

Bullying Parameters/Features

To identify the signs of bullying, we have taken multiple parameters under consideration through which we can filter-out real bullies from other behaviours which could be mistaken for bullying. Features like: rude behaviour, troubling authorities regularly, low attendance, use of foul language, low scores, drug use, explicit content etc. are used to filter the data. In the proposed model it is planned to integrate AI tools with school infrastructure and data like cameras with microphones, student portals containing community forums where students and teachers interacts, attendance of students, their score-cards etc. Data used for analysis and a complete infrastructural setup is discussed below in tables 1 and 2.

Table 1: How to identify a victim of bullying

Victim Parameter	Data Source
Show deteriorating grades	Academic report
Regular absenteeism	Attendance records
Being picked on, pushed, punched etc. (physically harassed)	CCTV images
Downfall of social skills	Registration for school events and extra-curricular groups
Suffering from learning or mental disabilities	Counsellor reports
Unwilling to go to school regularly	Attendance records
Random bruises, missing belongings or torn clothing	CCTV images
Prone to attacks of anxiety	School Medical Reports
Alone at lunch-breaks.	CCTV images
Experiences regular nightmares	Medical room report
Starts bullying younger or weaker kids to vent out the frustration	CCTV images

Table 2: Bullying perpetrator parameters and data sources

Perpetrator parameters	Data source
Rude behaviour with students as well as teachers	CCTV images, audio, school report
Low Grades	School report
Lacks empathy or guilt	School report
Feeling of entitlement because of being good in school, sports	School report

or belonging to a prominent family	
Short tempered and having emotional outbursts	CCTV images, audio, school report
Usually popular or among a big group	CCTV images, audio, school report
Regularly get into trouble with authority	School report

Integration with School Infrastructure:

Recordings of playground, common-area (like corridor, locker-rooms etc.) and classes will be taken along with their audios via cameras that are installed. More detailed information associated to students will be extracted from student portals. This will further help in understanding and analysing student behaviour and personality. This raw information is then reworked into structured data, which will supplement the learning algorithm in predictions and analysis. Based on this, appropriate action can be taken by school authorities. This method can also be inverted and be used to discern the victims of harassment.

Data Analysis:

Photos uploaded by the students on community forums along with acquired video footage from CCTV, split into images, and ran through the algorithm to identify: drugs, number of faces, anxiety attacks, crying, isolation, fighting, torn clothes, bruises, sleeping, smoking, hard drinks, gore, explicit and adult content. The image classifier uses CNN and trains the network from scratch and then converts the top layers of the network using VGG16.

Audios which are mapped to text using Google Cloud Speech API alongside comments at student community are further used to determine the following features: tone, amplitude and pitch of voice, language used (explicit or not), uppercase text, text length and sentiments, threatening statements, trolling, unpleasant comments and distasteful words. Other attributes like low-attendance, incompetent grades, enrolment in extracurricular activities, teacher-student interaction, frequency of councillor appointments and behaviour report by staff will also be considered. Data for these inputs is converted from physical form into digital form first.

Fig. 2. Sample images (in case of bullying) from dataset

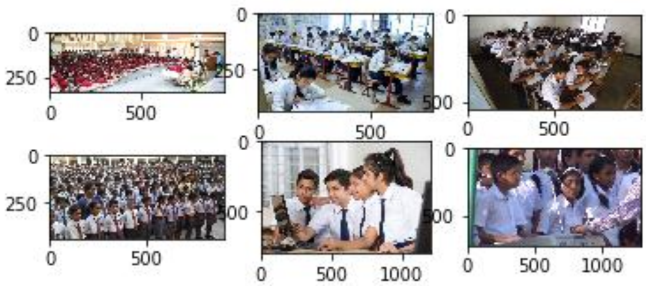

Fig. 3. Sample Images (non- bullying) from dataset

Techniques Involved

To implement our proposed model, Convolutional Neural Network for Natural Language Processing, along with Logistic Regression algorithms are used. NLP is broadly defined as the manipulation of natural language, e.g. speech to text.

Convolution Neural Network is a type of neural network (NN) that provides the capability to convert pixels into well structured data [5]. CNNs replicate the function of the frontal lobe of the human brain (cerebral cortex), which is responsible for processing audio visual stimulus in humans. To process images, CCTV footage is spliced into still images and then each frame is analysed to extract the crucial and important features which can be further analysed for more refined results.

Logistic Regression (sigmoid) function provides the capacity to differentiate between bullies and victims amongst other students from all collected features. It is also used within combination of other methodology to make a precise estimate. This method is suitable for most of the non-audio visual input data streams.

For audio analysis, we have mapped CCTV voice-output with Google Cloud Speech API which uses CNN and provides real-time streaming of speech recognition and conversion from audio to text. This lies within our CCTV and Google Cloud Storage for sentiment analysis to be done on it.

For text analysis, mentioned words are mapped with sentiment analysis dictionary (‘Liu and Hu opinion lexicon’ containing 6800 positive and negative words [4]). To extract feature from the text, multinomial Naïve Bayes is used. It helps in classification of bad words and rude comments from audio-to-text converted files as well as from student’s community portal including explicit and vulgar remarks, text length as well as usage of upper-case letters in community forums etc. All such attributes are then listed for feature extraction by the model.

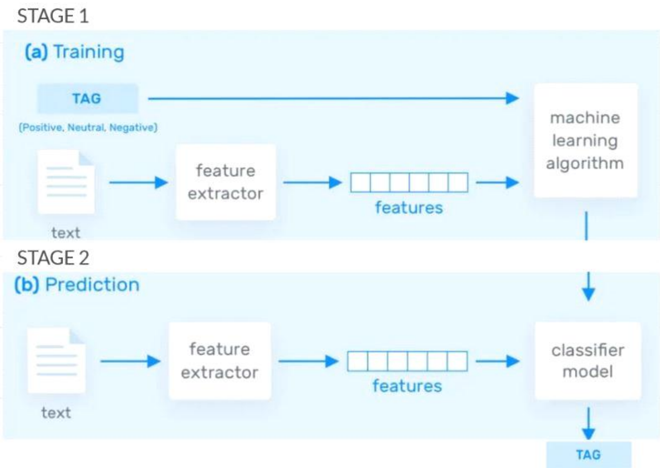

Fig. 4. Sentiment Analysis

Even data like medical records, attendance, grades and teachers remark from student portals is used and processed for feature extraction. Logistic regression modelling is used for this analysis.

For image classification, CNN is defined with multiple layers using sequential. Each layer is treated as an object that feeds data to next layer. This can be represented by a graph model allowing for interpretation of the working of the model. This feature map is then passed through an activation layer, called Rectified Linear Unit (RELU) [5]. RELU helps in increasing non-linear properties of our model. Thanks to it, our neural network will be able to learn more complex functions via Linear Regression only, which in turn, will help in distinguishing the traits of students in different images.

Implementation Infrastructure

Proposed Solution can easily be integrated and implemented in schools, colleges, institutes and other places prone to bullying. A data-pipeline can be built for fetching data at real-time with predictive analytical capabilities. Infrastructure for such a solution is a one-time investment and will provide a great benefit to the coming future making schools bullying free, a dream most people never thought would become a reality.

Detailed information and diagrammatic representation of Recommended Architecture is provided below:

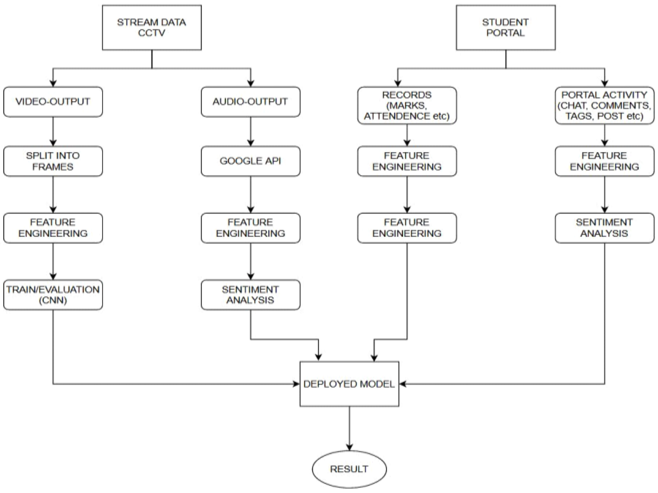

Fig. 5. Data-Flow/Infrastructure Architecture Flow diagram

As shown in the diagram, data streams from CCTV is split as audio and video media streams. Audio is then sent to Google speech-to-text API which is used for sentiment analysis. In case of video, clips are sliced down into multiple frames based on time and frame rates. These images are analysed to detect any signs of bullying or any other sort of unethical action via the use of CNN.

From student portals, web scraping is implemented to obtain information about posts, tags, comments, open-chats and for records which can be taken directly from the school database. All these attributes are fetched to be deployed into the model (on cloud premises) and results can be fetched remotely via internet.

Table 3: Feature Scaled Values

TEXT_LENGTH	TEXT_PERCENT_UPPERCASE	TEXT_SENTIMENT	TEXT_PERCENT_BAD_WORDS	AUDIO_TEXT_LENGTH	AUDIO_TEXT_SENTIMENT	AUDIO_PERCENT_BAD_WORDS	AUDIO_MUSIC_PERCENT	AUDIO_SILENCE_PERCENT	AUDIO_LOUDNESS	VISUAL_NUM_FACES	VISUAL_EXPLICIT	VISUAL_DRUG
5900	0.098983051	0.9998	0.018621974	124	0.6836	0	0.166666667	0	0.464696	0	0.17693	0.071883
2006	0.038384845	-0.9988	0.029255319	116	0	0	0	0	0.211718	0	6.97E-05	0.000127
476	0.071428571	0.9594	0.087912088	0	0	0	1	0	0.111151	0	0.19447	0.067911
2248	0.097864769	0.9839	0.029268293	0	0	0	0.142857143	0	0.263576	0	0.01258	0.023416
807	0.063197026	0.9595	0.037383178	0	0	0	0.142857143	0	0.18128	0	0.11153	0.078706
1451	0.101998622	0.5941	0.021097046	48	-0.25	0	1	0	0.071088	0	0.0067	0.001807
1545	0.102265372	0.9976	0.014440433	22	0	0	0.333333333	0	0.243846	1	0.60919	0.026073
930	0.033333333	0.9733	0.006666667	0	0	0	0.428571429	0	0.446988	0	0.00428	0.01015
417	0.254196643	-0.9047	0.095238095	0	0	0	0	0	0.254741	0	0.28213	0.009103
548	0.178832117	0.3423	0.064102564	10	0	0.5	1	0	-0.06896	0	0.30725	0.019809
575	0.07826087	0.8544	0.08	0	0	0	0	0	0.008551	0	0.00932	0.012283
616	0.167207792	-0.9812	0.091954023	44	0	0.125	0	0	0.26679	1	0.00086	0.001557
876	0.034246575	-0.4371	0.042553191	0	0	0	0.666666667	0	0.122666	1	0.00094	0.001101
529	0.094517958	0.9955	0	6	0	0	1	0	0.329987	0	0.00393	0.00235
1380	0.081884058	0.9668	0.011029412	25	0.34	0	0.333333333	0	0.19884	0	0.13785	0.023816
4706	0.113684658	-0.9966	0.055780933	0	0	0	0.857142857	0	0.158327	1	0.09014	0.039847
1582	0.087863464	0.9054	0.018248175	0	0	0	0.833333333	0	0.474464	0	0.04518	0.000836
1561	0.123638693	0.9252	0.013333333	13	0	0	0.833333333	0	0.042598	0	0.00072	0.000454
2544	0.055424528	-0.9791	0.042769857	0	0	0	1	0	-0.12871	0	1.54E-05	2.58E-05

Result

To analyse the accuracy in initial phases, labelled data was split into training and testing data in ratio of 7:3 respectively. At initial stages, with the gathered data our model is able to predict with an accuracy of 87 %.

The charts below depict cross-entropy loss which helps in measuring the performance of the classification model. The output represents a probability value between 0 and 1 [6] which shows our model's classification accuracy.

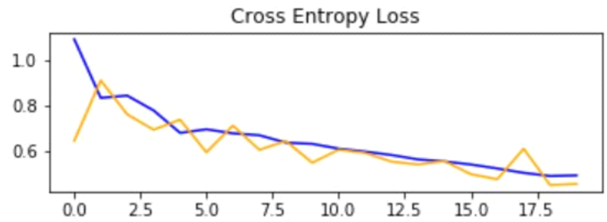

Fig. 6. Cross Entropy Loss (train dataset is represented by blue curve and test dataset is represented by orange curve)

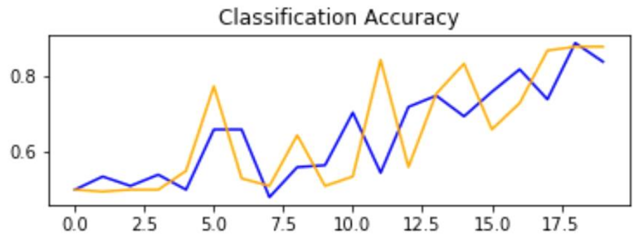

Fig. 7. Classification Accuracy (train dataset is represented by blue curve and test dataset is represented by orange curve)

	BULLY	VICTIM	NORMAL
BULLY	470	42	18
VICTIM	38	1131	51
NORMAL	50	102	399

Fig. 8. Confusion Matrix (This matrix is plotted on the labelled data having sample space of 2000 to check the accuracy of the deployed model)

Improvements

Future Improvements can include real-time identification of bullies through facial recognition. Sending of alerts to management in case of any real time bullying incident. The model can also be used as a measure to check for depressed victims so as to avoid suicide and self-harm tendencies and help in countering it.

Conclusion

The proof of concept solution developed by the authors was successful in identifying the bullying situation and hence the bully and the victim. The output was more accurate for audio-visual inputs. For the unstructured data, the accuracy of prediction will have to be improved by inter linking of the outputs with the audio visual parameters. Overall it was a successful proof of concept.

Acknowledgment

We would like to thank Mr. Kirit, the CEO of Gazelle Information Technologies PVT LTD, for his expert advice and a supply of required resources for the implementation of this project.

References:

1. National Academies of Sciences, Engineering, and Medicine. 2016. Preventing Bullying Through Science, Policy, and Practice. Washington, DC: The National Academies Press. https://doi.org/10.17226/23482.
2. D. Poeter. (2011) Study: A Quarter of Parents Say Their Child Involved in Cyberbullying. pcmag.com. [Online]. Available: http://www.pcmag.com/article2/0,2817,2388540,00.asp.
3. StopBullying.gov. (2020). What Is Bullying. [online] Available at: https://www.stopbullying.gov/bullying/what-is-bullying [Accessed 28 Jan. 2020].
4. Liu, B., 2020. Opinion Mining, Sentiment Analysis, Opinion Extraction. [online] Cs.uic.edu. Available at: <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon> [Accessed 16 June 2020].
5. WildML. (2020). Understanding Convolutional Neural Networks for NLP. [online] Available at: http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/ [Accessed 28 Jan. 2020].
6. Brownlee, J. (2020). How to Classify Photos [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-to-classify-photos-of-dogs-and-cats/ [Accessed 28 Jan. 2020].

Biographies:

Lalit Kumar is an associate consultant with Gazelle Information Technologies, New Delhi. He holds a bachelors in technology with special interests in AI and using AI for practical applications. He has built practical solutions for businesses to predict incidents in the supply chain using various algorithms.

Palash Goyal is currently a student of class XII at Mount Carmel School, Sector 23, Dwarka, New Delhi. He started coding at an early age and is currently working on Python, Automation and Image Processing. It was his personal experience at one of his previous schools, that led to the birth of this idea last year, which he converted into reality with the help of other authors.

Karan Malik is a 2nd year student of B. Tech Computer Science at University School of Information, Communication and Technology (USICT), Indraprastha University, New Delhi. His area of interests includes machine learning, image processing and natural Language Processing.

Rishav Kumar is currently a student of class XII at Mount Carmel School, Sector 23, Dwarka, New Delhi. His areas of interest include AI and internet of things.

.