WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI Image Classification for Cyberbullying Using Mobilenet V2 Transfer Learning

[1]Mr. R. Sathish Kumar, MCA, M. Phil., [2]T. Sangeetha

[1]Asst. Prof., Department of MCA, Krishnasamy College of Engineering & Technology
[2]Student, Department of MCA, Krishnasamy College of Engineering & Technology

ABSTRACT

The impact of cyberbullying on the lives of victims is immeasurable. Because how the person deals with it is very subjective. The message may be brutal to victims, however to others it is able to be normal. Ambiguities in cyberbullying messages create an incredible mission in finding dependable content material. Some studies has been said to deal with this query with a textual content-primarily based method. However, picture-based detection of cyberbullying is receiving less attention. This task aims to increase a model that allows prevent image-primarily based cyberbullying issues from being posted on social media. This task advocates an automatic version based totally on switch mastering to discover a cyber-symbol photo from a malicious social platform. Transfer gaining knowledge of fashions can extract hidden contextual functions from cyberbullying messages. Our test includes two photograph datasets (i.e. Composed of cyberbullying and non-bullying images). The datasets may be beneficial for destiny researchers to extend the studies. It is hard to find the pleasant and suitable version for detecting bullying images, so test with both DL and transfer gaining knowledge of models to locate the first- class model. The experimental effects confirmed that switch gaining knowledge of models are the excellent desire for predicting image-based totally cyberbullying messages.

1. INTRODUCTION

With the development of Internet 2.0, there was a giant trade in social verbal exchange, and relationships and friendships were restructured. Almsgiving spaces of time on line and on one of a kind social structures, similarly to all the advantages that they bring, their presence additionally makes them susceptible to threats and social offenders which include cyberbullying.

Cyberbullying ought to be understood and addressed from exclusive angles. Automatic detection and prevention of those incidents can considerably assist resolve this hassle. Tools have already been evolved that could file incidents of bullying and applications that try to provide help to victims. Additionally, most on line platforms which might be generally utilized by teenagers have protection facilities, as an example, the YouTube Safety Center and Twitter Safety and Security, which offer user help and reveal verbal exchange. A lot of studies has additionally been finished on automaton detection and cyberbullying prevention, which we will discuss in more element inside the subsequent phase, but this problem is still a long way from finding a definitive answer and similarly (DNN) based fashions have additionally been applied to cyberbullying detection.

With the increasing demand for online social media, a new form of bullying has emerged. It is defined as an aggressive, intentional act carried out by a group of individuals, using electronic forms of contact repeatedly and over time against a victim who cannot easily defend him or herself [1].

The form of cyberbullying can vary from text to photos and videos by circulating false rumors and disclosure of personal information that are directed to harm and discredit the victims [2]. The inability to identify cyberbully behavior embellishes the threatening and intimidating nature of the intrusions for victims. In addition, the consequential effects (introduced later) of cyberbullying are unlimited given the virtual world within which cyberbullies function, as they can engage in these activities in a relentless fashion without regard to time [3].

The primary victims of cyberbullying are teenagers. From Pew research center in 2014, 73% of internet user have witnessed online harassment and 40% have personally experienced it. In a survey where five types of bullies are included, offensive name-calling and purposeful embarrassment were the most common types of harassment people witnessed.

There are also a variety of harms that cyberbullying can do to people, especially teenagers. According to stopbullying.gov—the US government website for bullying prevention, teens who are cyberbullied are more likely to: use alcohol and drugs, skip school, experience in-person bullying, be unwilling to attend school, receive poor grades, have lower self-esteem and have more health problems. More severely, Hinduja, et al. conducted a study and shows that cyberbullying victims were 1.9 times more likely and cyberbullying offenders were 1.5 times more likely to have attempted suicide than those who were not cyberbullying victims or offenders.

Given the severity of cyberbullying, actions have to be taken to detect and prevent it. Previous work has been focusing on the detection of cyberbullying after it already happened. The purpose of that is to recognize and stop the cyberbullying offenders in the virtual world. However, once the

bullying words are created, they are hard to be taken back or intercepted due to their fast and wide spreading nature. Therefore, the best way to efficiently prevent cyberbullying from happening is to stop providing bullies with bulliable contents.

The purpose of this thesis is to find a prediction method to advise social media sites users about what and what not to post in their accounts in order to avoid being bullied. More specifically, we focus on predicting bullyable images and photos on social media, which means that the images that we are researching are not those posted in purpose of harassing people but those that are harassed. They may receive bullying comments or be modified intentionally to embarrass the poster and other viewers etc. In this thesis, we focus on those bullied via textual comments.

Bullyable images refer to the images that are possible attacking targets for online bullies. It is hard to generalize the content of those images, since bullies have their own preferences on their targets. Yet, we can still semantically differentiate bullyable images from bullying images, which is important before we continue. Bullying images are those that initialize embarrassing or painful feelings to viewers while the opposite type, innocent images, do not intend to do that. Although they are different in content, they could both become targets for online bullies. Our aim is to find an efficient way to recognize bullyable images, bullying or not, and suggest users not post them.

We used several machine learning algorithms to learn the pattern of bullied images (training set of images) and detect the bully ability of other images (test set of images), while the images are all labeled bullied/non-bullied manually. We made it a two-step job to decide if an image is bullied given its comments: first, check if there is(are) bullying comment(s) among all; second, check if those bullying comments are targeting at the content of the image. The steps are followed by the Internet crowd to perform the label work for us.

The data set of this thesis is crawled from Instagram, a social media site especially popular among teenagers. According to a survey by Pew Research Center, Instagram, being the second most popular social media site among adolescents, is used by 52% of teens. And 40% of them claim Instagram is their most used social media website. Instagram also has the most significant growth (9%) in overall user figures among all the social media sites[5]. Instagram is an online photo and video sharing social media website. Also, it confines photo posts to a square shape with a relatively high resolution. Those characteristics make Instagram an ideal data source for image-oriented research studies. Also, the popularity of it gives more image posts and comments for us to learn the pattern of cyberbullying.

2. LITERATURE SURVEY

As the negative influence caused by cyberbullying is increasing, an increasing number of studies are dedicated to dealing with, mainly detection of, cyberbullying. There are non- technical works concentrating on giving definitions, reporting status quo and understanding the problem of cyberbullying [6-8]. Those studies give directions for our detection work. There are still not many studies dedicated to automatically detect cyberbullying but the number is increasing [9-14].

The causes of cyberbullying and its prevalence, especially for children and young adults, have been extensively studied in social sciences research [6]. In terms of its impact, empirical studies have demonstrated a link between suicidal ideation among adolescents and experiences with cyberbullying [7]. The characteristic profiles of offenders and victims in cyberbullying are presented in [8]. This paper also discusses the possible strategies of prevention and intervention. Those studies enlighten us on the scope and spread the awareness of the problem.

Most technical studies concentrate on text analysis. Yin, et al. added sentiment and contextual features to baseline text mining system thus enhanced its performance in online harassment detection [9]. The study provides an accuracy of up to 50% on the dataset of Kongregate, Slashdot and Myspace. A similar work reached an accuracy of 80% in the detection of bullying comments on YouTube through textual context features [10]. In [11], Reynolds et. al. studied the language patterns of online bullies and victims on a small data set from Formspring and developed rule-based classification systems to identify the bullies. Their model is powerful enough to distinguish 78.5% of the bullying posts. To further explore the capability of language features in cyberbullying detection , Xu, et al. established a new sentence-level filtering model that semantically eliminates bullying words in texts by utilizing grammatical relations among words [12], whose results are very close to manual filtering, 90%, from experiments on YouTube dataset.

Besides pure text analysis, other techniques are involved to assist the detection. Bayzick, et al. involved a truth set while building a rule-based classification model to detect the presence of cyberbullying in online conversation [13].

3. PROPOSED SYSTEM

System Modules :

• Data collection

• Data preprocessing

• Performance evaluation

• Flask framework

Module Description :

1. Data Collection:

The information was gathered to allow the model to identify every day and cyberbullying cases. For example, if obscene photographs with human faces are present, the example of a commonplace human face is about as bullying. To avoid this example, a sufficient range of records instances were added in every category.

Negative Bulling

Positive Bulling

2. Data Preprocessing:

Convolutional Neural Networks (CNs) can do wonders if there may be enough statistics. However, the choice of the right education information set of all of the characteristics which are essential for training is a hard undertaking. If the consumer does now not have enough, the network can load schooling information. Realistic photographs contain distinct sizes, positions, zoom, lighting, noise, and so forth.

For a robust community of these commonly encountered elements, the Data Augmentation method is used. By rotating input pictures to one-of-a-kind angles, flipping pictures via one of a kind axes, or rotating/cropping pix, the community will come across these phenomena in practice.

Using Keras, there may be a handy institution of arguments in the "Image Data Generator" that allows picture augmentation. By flipping the photo or vertically, we can absolutely rearrange the factors, however the capability of the system is retained. That may be performed in numpy.

Rotation is an increment this is typically accomplished at angles of 90 levels, but can also be done at smaller or minute angles if extra-large facts is needed. In fact, the history colour of the rotation is normally fixed so that it blends into the historical past. Otherwise, the model may also count on that the trade course is separate. This works excellent while the difficulty is the same in all turned around pix.

Keras Image Data Generator class has two most important parameters width_shift_range and height_shift_range and based totally on those values the photo factors will shift both left or up and down.

• At the time of schooling, it best includes studying one repressors or one binary classifier according to class. This being the case, it's far essential to label the multi-genres into binaries (they belong or do not belong to a genre). Label Binarizer makes this method easy with the remodel technique.

- Convert to array and normalize to interval of [0,1]
- Fetch images and class labels from files
- Arrange format as per keras
- Resize as per model
- Append image
- Append class label
- Make labels into categories - either 0 or 1, for our model
- Label binarizer

BLOCK DIAGRAM

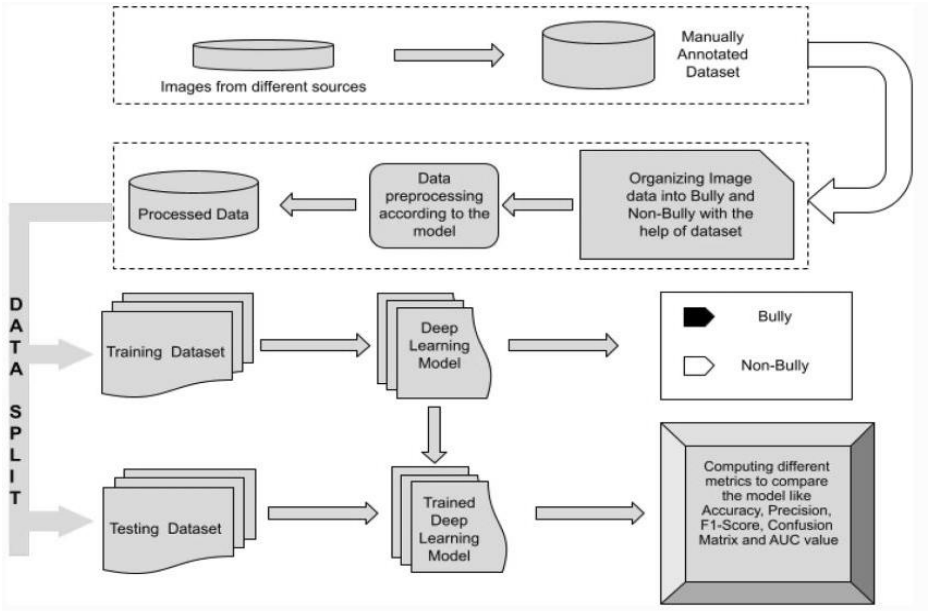

Fig. 1: System Architecture

3. PERFORMANCE EVALUATION

Classification Accuracy

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

It works well only if there are equal number of samples belonging to each class

Logarithmic Loss

Logarithmic Loss or Log Loss, works by penalising the false classifications. It works well for multi-class classification. When working with Log Loss, the classifier must assign probability to each class for all the samples. Suppose, there are N samples belonging to M classes, then the Log Loss is calculated as below**:**

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

4. FLASK FRAMEWORK

Routing:

In Flask, routing refers to mapping URLs to specific functions or views. You define routes using decorators (@app.route()) to associate URLs with corresponding functions in your code. For example, @app.route('/home') associates the /home URL with a specific function.

Views and Templates:

Views are Python functions that are associated with specific routes. When a user accesses a particular URL, Flask invokes the corresponding view function. Views can return rendered templates, which are HTML files containing dynamic content. Flask uses the Jinja2 templating engine to render these templates with data.

Request Handling:

When a user makes a request to the Flask application, Flask handles the request and passes it to the appropriate view function based on the URL and HTTP method (GET, POST, etc.). The view function processes the request, accesses the data submitted by the user, performs the necessary operations, and prepares a response.

Response Handling:

Flask allows you to define the response to be sent back to the user. This can include rendering templates, returning JSON data, redirecting to another URL, or returning static files such as images or CSS files

Integration with Cyberbullying Detection:

In the context of cyberbullying detection, Flask can be used to integrate the cyberbullying detection model with the web interface. When a user uploads an image or inputs text, Flask can pass this data to the cyberbullying detection model for analysis. The result of the analysis can then be displayed to the user through the Flask-based web interface.

Deployment:

Flask applications can be deployed on various web servers or platforms, including local development servers, cloud platforms, or production servers. Flask provides a built-in development server for testing, but for production, it is recommended to use a more robust web server such as Nginx or Apache.

4. CONCLUSION

Compound problems like cyberbullying that have more than one built in questions, are difficult to investigate with a normal gadget. In unique, the put up-detection of image-based social cyberbullying is a difficult mission. This research investigated deep learning and transfer mastering frameworks to discover the quality fit version to be expecting cyberbullying messages on social media structures primarily based on photo. Mobile switch mastering fashions done higher and more accurate predictions. Therefore, it could be concluded that the proposed device detects maximum photo-based totally cyberbullying messages.

The boundaries of the proposed model are: (i) it does not recollect the detection of textual cyberbullying, which means that textual content-simplest messages are not a part of this studies, (ii) the mixture of photos with textual content in cyberbullying messages. However, this have a look at is limited to photograph-based cyberbullying detection. Therefore, the future scope of this studies is continually open to dialogue because the sub-varied problems. A piece of textual content can be viewed alongside an photograph to locate extra cyberbullying posts on social structures.

5. FUTURE ENHANCEMENT

Multimodal Analysis:

Expanding the system to incorporate multimodal analysis can enhance its effectiveness. Besides analyzing images, incorporating textual analysis and contextual information from social media posts can provide a more comprehensive understanding of cyber bullying incidents. By combining image analysis with natural language processing techniques, the system can detect and classify cyber bullying instances more accurately.

Fine-grained Classification:

Refining the model to perform fine-grained classification can provide deeper insights into the nature and severity of cyber bullying. Instead of simply categorizing posts as cyber bullying or non-cyber bullying, the system can be enhanced to classify different types of cyber bullying behaviors, such as harassment, hate speech, or threats. This level of granularity can assist in tailoring appropriate interventions and support strategies.

Adversarial Attacks Detection:

Considering the possibility of adversarial attacks is crucial for improving the robustness of the system. Adversarial attacks involve intentionally modifying images to deceive the model and evade detection. Developing techniques to detect and mitigate such attacks can enhance the system's reliability and prevent malicious manipulation of images to bypass detection.

Real-time Monitoring:

Enabling real-time monitoring of social media platforms for image-based cyber bullying can provide timely intervention and support. The system can be enhanced to continuously analyze newly posted images, detect potential cyber bullying content, and alert platform administrators or relevant authorities. Real-time monitoring allows for swift response and intervention to prevent harm and provide assistance to victims.

6. REFERENCE

[1] Chollet F (2017) Xception: Deep gaining knowledge of through separable deep convolutions. In Proceedings of the IEEE Conference on Computer Vision and Recognition, pp. 1251-1258

[2] He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Recognition, pp. 770- 778

[3] Roy PK, Ahmad Z, Singh JP, Alryalat MAA, Rana NP, Dwivedi YK (2018) Finding and Ranking High Quality Answers on Community Questions Answers Sites. Global J Flex System Management 19(1):53 -68

[4] Bhat S, Koundal D (2021) Fusion of multifocal pics the usage of neutrosophic-based wavelet remodel. Soft Computing App 106:107307

[5] Kumari K, Singh JP, Dwivedi YK, Rana NP (2021) Multimodal Aggression Identification Using Convolutional Neural Network and Particle Binary Optimization. Future Gener Computing Syst 118: 187-197.

[6] Aggarwal S, Gupta S, Alhudhaif A, Koundal D, Gupta R, Polat K (2021) Automated detection of COVID-19 in chest X-ray photos using exact architectural studies. Expert System 39:e12749

[7] Modha, S.; Majumder, P.; Mandl, T.; Mandalia, C. Detecting and Visualizing Hate Speech in Social Media: Cyber Security for Surveillance. A completely skilled device. App. 2020, 161, 113725. [Cross-reference]

[8] Dinakar, K.; Reichart, R.; Lieberman, H. A textual version for detecting cyberbullying. In Proceedings of the AAAI International Conference on Web and Social Media, Atlanta, GA, USA, June 6-9, 2012.

[9] Dadvar, M.; Jong, F.D.; Ordelman, R.; Trieschnigg, D. Improving Cyberbullying Detection Using Gender Information. In Excursus host guide titled Journal of the twelfth Dutch-Belgian information research workshop (DIR 2012), Ghent, Belgium, on February 24, 2012.

[10] Kontostathis, A.; Reynolds, K.; Garron, A.; Edwards,L. Cyberbullying Detection: Terms and Techniques Issues. In Proceedings of thefifth Annual ACM, Internet of Sciences Conference, online, 2 May 2013; p. 195-2