JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 8, ISSUE 04, 2021

Predicting Cyberbullying on social media

Priyabrata Nayak[1], Malla Reddy Meka[2], Akshay Prasad Satapathy[1]

[1]Assistant Professor,[2]Associate Professor,[1,2]Dept. of CSE

[1,2]Gandhi Institute for Technology, Bhubaneshwar, India

Abstract

Cyberbullying is the use of Information and Communication Technology (ICT) by individuals to humiliate, tease, embarrass, taunt, defame and disparage a target without any face-to-face contact. Social media is the "virtual playground" used by bullies with the upsurge of social networking sites such as Facebook, Instagram, YouTube, Twitter etc. It is critical to implement models and systems for automatic detection and resolution of bullying content available online as the ramifications can lead to a societal epidemic. This research proffers a novel hybrid model for Cyberbullying detection in three different modalities of social data, namely, textual, and info-graphic (text embedded along with an image). The architecture consists of a Deep Learning convolution neural network (DLCNN) for predicting the textual bullying content. The info-graphic content is discretized by separating text from the image using Google Lens of Google Photos App. The processing of textual and visual components is carried out using the hybrid architecture and a Boolean system with a logical OR operation is augmented to the architecture which validates and categorizes the output on the basis of text and image bullying truth value. The model achieves a prediction accuracy of 98% which is acquired after performing tuning of different hyper-parameters. The simulation results show that the proposed method gives the better accuracy compared to the state of art approaches.

1. Introduction

Through Cyberbullying, an individual or victim can be humiliated or hurt before whole network on the web[1]. It bothers the psychological and physical condition of an individual because of which an expansive number of suicides and discouragement cases happen. These days, Cyberbullying, through images or memes are extremely common. Pornographic images or pictures with oppressive, mean or defamatory remarks are being presented on one's profile in order to bully them. Protective measures must be taken so as to control this. Thus, in our work we have made an automated framework which will, in addition to detection of bullying for text [2], also recognizes the bullying in the pictures. There can be a straightforward picture with no content on it or there might be such sort of picture also where some content is embedded with that picture. More recently, as memes and GIFs [3] dominate the social feeds; typo-graphic and info-graphic visual content has become a considerable element of social data. Thus, cyber bullying, through varied content modalities is very common. Researchers worldwide have been trying to develop new ways to detect cyber bullying, manage it and reduce its prevalence on social media. Advanced analytical methods and computational models for efficient processing, analysis and modeling for detecting such bitter, taunting, abusive or negative content in images, memes or text messages are imperative. Social media specificity, topic dependence and variety in hand-crafted features currently define the bottlenecks in detecting online bullying posts [4]. State-of-the-art results are achieved by deep learning methods on some specific language problems by using the capabilities of hierarchical learning and generalization [5]. Pertinent studies report the use of deep learning models like CNN[6], RNN[]7 and semantic image features for bullying content detection by analyzing textual, image based and user features [8].But most of the research on online cyber-aggression, harassment detection and toxicity has been limited to text-based analytics [9]. Visual analysis of images have also been reported by few related studies for determining bullying

ISSN- 2394-5125 VOL 8, ISSUE 04, 2021

content [10-12] but the domain of visual text which combines both text and image has been least explored in literature. The combination can be observed in two variants: typo-graphic (artistic way of text representation) or info-graphic (text embedded along with an image). This research work puts forward a hybrid deep learning model for bullying content prediction, where the content, c ε {text, image, info-graphic}.The primary contribution of the work is that unlike previous models which are mono-modal dealing with a single type of content modality.

The major contributions of the paper as follows:

- The proposed method is capable of detecting the bullying data from the both Textual and info-graphic images respectively.
- The overall purpose of this analysis is therefore to evaluate the performance of standard DLCNN classification algorithms by using SIFT features for bullying language identification.

This paper is summarized follows through as: In Section 2, literature review for cloud data security with the comparison of methodology with defining problem, implication, merits and demerits. Section 3 gives the detailed information about the proposed methodology. Section 4 discusses about the results analysis and finally Section 5, concluded the summarization of whole paper.

2. Literature Survey

Twitter is an online social networking platform, also referred to as a microblogging site, where users can share information via messages up to 140 characters. These short messages are called 'tweets'. Twitter allows users to tweet through its website or through its applications developed for various external compatible devices. In most countries, users can also use SMS services to tweet. Tweets can be read by anyone unless the users restrict access strictly to their followers. When a user subscribes to another user account, the subscription is termed 'following' and the subscriber is called the 'follower'. Twitter, as a social networking platform, spins around the term 'followers'. For example, if user A is following user B, user A as a follower gets access to read and retweet user B's tweets. Out of all the tweets generated on Twitter, roughly 40% of tweets are conversational tweets (Kelly, 2009). Users make use of hashtags to tweet about trending topical information. Similarly, users make use of '@' followed by a username, for example '@username' to post a tweet mention or reply to another user.

One thing that makes it easy for bullies to harass someone online is that they can retain their anonymity by creating fake accounts to bully someone. Due to the functionality of 'hashtags' and '@username', victims are more vulnerable to direct online attacks. In addition, the victims are totally exposed, as their followers can witness the entire cyberbullying episodes. Twitter provides a system to reduce cyberbullying, but unfortunately it is not effective. Twitter has a 'report abuse' form that users affected by cyberbullying must fill out if they wish some action needs to be taken. Twitter needs a more intelligent system to detect cyberbullying, which is more efficient in detecting cyberbullying tweets on Twitter.

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 nontechnical works concentrating on giving definitions, reporting status quo and understanding the problem of cyberbullying [13]. Those studies give directions for our detection work. There are still not many studies dedicated to automatically detect cyberbullying but the number is increasing [14].The causes of cyberbullying and its prevalence, especially for children and young adults, have been extensively studied in social sciences research [15]. In terms of its impact, empirical studies have demonstrated a link between suicidal ideation among adolescents and experiences with cyberbullying [16]. The characteristic profiles of offenders and victims in cyberbullying are presented in [17]. This paper also

ISSN- 2394-5125 VOL 8, ISSUE 04, 2021

discusses the possible strategies of prevention and intervention. Those studies enlighten us on the scope and spread the awareness of the problem.

The research introduced above all focuses on the detection of bullies, but in this work, our target is different. As mentioned before, we attempt to predict if an image posted on social media would become a target for the bullies. Therefore, instead of spotting the bullies, we focus on preventing people from feeding the bullies. Moreover, instead of text information, we utilize image features to build classification model. The expected outcome is a model that can warn social network site users before they post any images that are highly "bullyable". However, as far as we know, there is not any other work that is based on image features in cyberbullying detection. In this work, we explore the potential of existing image feature extracting methods in other area and make them function in our work.

3. Proposed Method

The proposed deep classification model reinforces the strengths of deep learning nets in combination to machine learning to deal with different modalities of data in online social media content. The proposed DLCNN model consists of four modules namely, text analytics module, info-graphic analytics module, discretion module and decision module. The basic architecture of the work done has been clarified beneath as shown in Figure 1. The steps included can be comprehended as

- Analyzing the sort of information.
- Passing it to the separate module for processing.
- Decision module is utilized to analyze the outcome.

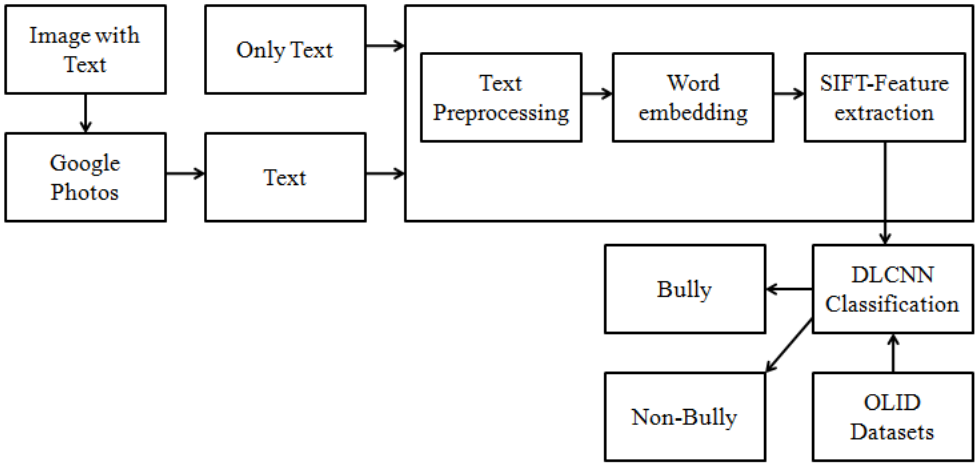

Figure 1: Proper flow of proposed model

Analyzing the sort of information includes checking whether the input is just text or it is a picture or it is a picture with text embedded on it. This is vital in light of the fact that once we have investigated this then we can perform further handling in the respective modules. On the off chance that the input is as just text, at that point we will perform pre-processing of text, extract the features, create the feature vector and after that utilization DLCNN is used for performing the task of classification

If the input is the image with text embedded on it on it, at that point an additional step will be included to separate that text from the picture, which we are doing utilizing Google Photos as a tool. When we have the text separated from the image, we can utilize the means utilized for performing text analysis and for the picture we will utilize the image handling steps. The result of those two

ISSN- 2394-5125 VOL 8, ISSUE 04, 2021

modules will be encouraged as contributions to a Boolean framework that will at that point shows the outcome whether it is a bully or not.

When one of the input to Boolean framework isn't accessible like we have only text or in the event that we have picture just, at that point that input to Boolean framework will be unfilled or false since we are utilizing an OR operation in the Boolean framework, if the text or image is a sort of bully, it will get identified.

Preprocessing: Fortunately, Twitter provides an official application program interface (API), which makes it convenient for us to download the data we needed. However, a proper acquisition strategy was still needed in consideration of the existence of duplicate data, foreign language text and uninformative comments. And to purify the texts for labeling and machine classifying, we filtered the emojis and less informative comments.

Word Embedding: The pre-processed posts are input into the embedding layer. The feature representation and extraction in the DLCNN is learned in a hierarchical way using word embeddings making it distinctive and better than the lexical or syntactic feature extraction. The embedding layer thus uses GloVeto build word embeddings and the model learn geometrical encodings (vectors) of words in each post. We run our model on top of GloVe word embedding using 100 dimension representation of word. We train the system to learn the vectors for each word (which would be represented as one hot vector initially), thus we convert each word to a vector of integers of 100 dimensions and therefore we have a comment matrix of size equals to number of words in the vocabulary multiplied by 100. Now our text data is in the form of numerical data that can further be used for performing convolutions.

Feature extraction: In this section we present the features involved in classification of images into the classes “bullied” and “non-bullied”. The purpose was to detect images being bullied based on image content features. In this work, we used edge-direction coherence vector, color histogram, Scale-invariant feature transform (SIFT) and face (number and ratio) features to describe the images. Besides those, caption and user information were also involved as auxiliary features to help the classification.SIFT is a common technique to perform object recognition on images. By using SIFT; we explore the similarity of objects contained in bullied images and the similarity of objects contained in non-bullied images relatively. In SIFT 128-dimension descriptors are formed for points that are considered interesting in an image by the algorithm. There are multiple methods to design features for images that are built upon these descriptors. One popular technique is to quantize the interesting points into “words” and build a dictionary of visual words. With that we can create sparse vectors with predetermined length for each image that encode the number of appearances of each type of visual term. In this work study, we find interesting points in each image and perform k-mean clustering to learn a codebook for vector quantization from the whole image set being trained and tested. For the size of codebook, we tried 500, 1000 and 1500.

CNN for Performing Text Analytics

The concepts used in Convolution Neural Network consist of various terminologies which are briefly defined as: Convolution:

- Convolution: Applying filter to a fixed size window is the task of convolution operation
- Convolution Filter: It is also known as convolution kernel. It is basically a matrix that is utilized for performing convolution operation.
- Pooling: It is the process of combining the vectors obtained as a result of various convolution windows into a vector single one dimension.

ISSN- 2394-5125 VOL 8, ISSUE 04, 2021

- Feature maps: The significance of number of feature maps is that it directly controls capacity and is dependent on count of available examples and complexity of task.

The deep neural architecture for text analytics is shown in Figure 2. The figure describes the process of applying CNN over text in order to perform the task of classification. This demonstrates a step by step explanation of every process involved during the application of CMNN algorithm.For the application of DLCNN over text, its components should be encoded before it is given as input to DLCNN. For this, a vocabulary is used which is formed as an index having words that appear in the posts/comments. Each word is mapped to an integer between 1 and vocabulary length. An example text encoding using vocabulary is shown in figure 3.

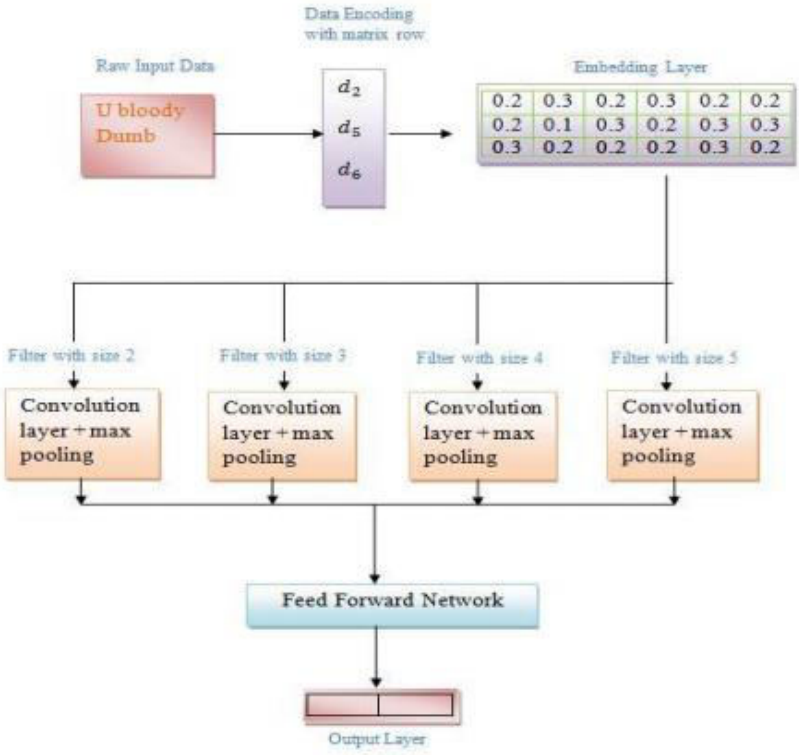

Figure 2: DLCNN for text classification process

Padding is used to maintain the fixed input dimensionality feature of DLCNN, in which zeros are filled in the matrix to get the maximum length amongst all comments in dimensionality. In the next step the encoded texts are transformed into matrices where each row represents one word. The constructed matrices pass through the embedding layer where each word (row) is converted into a lower-dimensional representation by a dense vector. Next step is to perform convolution, one way to think of convolution is that we're sliding the filter over the input text. For each position of the filter, we are multiplying the overlapping values of the filter and text together, and add up the results. This sum of products will be the value of output at the point in the input text where the filter is centered.

The process then continues following the generic CNN model like passing this obtained n-dimension matrix to the feed forward network and finally result is obtained by the output layer.A fully connected neural network is a feed forward network that will have the feature vector of n dimension obtained after concatenating every obtained by the application of n filters. Now we train the network using back-propagation algorithm. Gradients are back propagated and when we reach at the convergence we finally stop the algorithm. A softmax function is used to classify the post as bully (+1) or non-bully (-1).

ISSN- 2394-5125 VOL 8, ISSUE 04, 2021

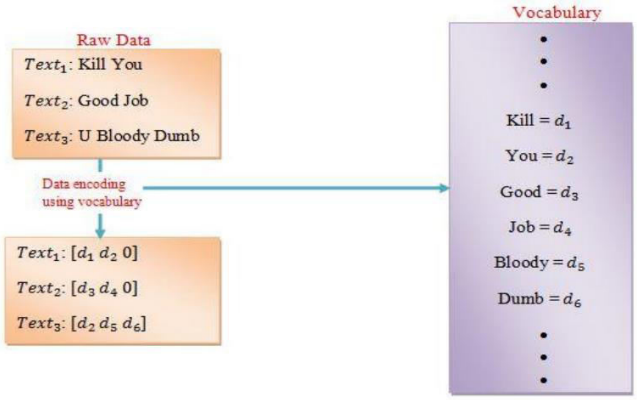

Figure 3: Text Encoding using Vocabulary

4. Results and analysis

4.1. Data Set:

The OLID dataset by Zampieri et al., 2019 is obtained via Twitter APIs using filters on certain keywords set. This data is then annotated using human annotators. To annotate one tweet, at least two out of three annotators must provide it the same label. By application of various sampling techniques, the distribution of offensive tweets is maintained at about 30% of the entire dataset.The data used here is in generally three forms with percentages shown below. The dataset prepared for experiments contains 10000 data out of which 55% data is in the form of text, 21% is in image form and there are 24% of data that contains image embedded with text. Table 1 below shows actual distribution of data in numbers. The dataset is prepared using the social media sites like Youtube, Instagram and Twitter. Comments and posts from these sites have been gathered and used for experiments.

Table 1: Categorization of data used in input

Type of modality	Number of instances	
	Bully	Non bully
Image only	1260	840
Text only	3300	2200
Info-graphic	1440	960

4.2 Performance evaluation

In our overall model the text used is the text without removal of emoticons during preprocessing. Confusion matrices for all type of modalities are shown in Table 2.

Table 2: Confusion matrix for textual modality

Actual classification TEXT	Predicted classification			
	Bully	Non bully	Precision	Recall
Bullying	5100	900	84.25%	91.83%
Non-bullying	1280	2720	85.84%	74.25%

For the textual module, five classifiers, namely, Naïve Bayesian (NB), Random Forest (RF), Support vector machine (SVM), K-nearest neighbor (KNN) and Sequential Minimal Optimization (SMO) were compared with CNN. CNN achieved the highest accuracy of 78.2%. Two classifiers, namely, KNN and NB were compared with SVM using the BoVW features for the image analytics moduleand

ISSN- 2394-5125 VOL 8, ISSUE 04, 2021

it was observed that SVM outperformed both the other classifiers. Comparative analysis of the classification algorithms used for discrete textual and visual modalities is given in Table 3.

Table 3: Classification Results for Textual, Visual and Info-graphic modalities

Classifier	Accuracy	Precision	Recall
NB[14]	69.6%	79.36%	66.66%
RF[15]	69.1%	77.50%	68.33%
SVM [17]	67.8%	76.62%	66.66%
DLCNN	98.2%	89.94%	85%

Here we have explained the setting of various parameters that has been used for performing the experiments. We have defined the proper distribution of the data as in what proportion the modalities are used in our model. Further we have analyzed our model individually for each type of modality and analyzed the results. The results are also compared by using different classification algorithms like naïve bayes, SVM, KNN etc. and observed CNN gave the best results for text modality among all the methods. The best result obtained after setting all the hyper-parameters is 98.2% for our model.

According to the results, the contribution of general image features to the classification work is limited. One of the possible explanations is that the patterns of bullying are beyond merely image characteristics. On social media sites, a viewer comment under a post knowing information a lot more than just the post itself which is hard to simulate in a classification model. While trying different techniques to classify the data set and enhance the performance, we discovered a big diversity of the image content and patterns of bullying. Due to the popularity of Instagram, Facebook or Youtube users post photos or comments with different purposes. Some post to promote products, some post to report news, some are organizations and post to gain popularity among viewers and some are common individuals who post to share experiences of their life. The bully might be intrigued by the identity of a user posting the comment in form of text, photo, or the content of images that the people who comment have strong sentiment about. These factors might require specific common sense knowledge to be recognized, which is sometimes hard for others without it to see and increase the difficulty of this classification problem.

5. CONCLUSION

Social media and the internet have opened up new forms of both empowerment and oppression. Meaningful engagement has transformed into a detrimental avenue where individuals are often vulnerable targets to online ridiculing. A predictive model to detect this Cyberbullying in online content is imperative and this research proffered a prototype model for the same. The uniqueness of the proposed hybrid deep learning model, DLCNN is that it deals with different modalities of content, namely, textual, and info-graphic (text with image). The results have been evaluated and compared with various baselines and it is observed the proposed model gives superlative performance accuracy. The limitations of the model arise from the characteristics of real-time social data which are inherently “high-dimensional”, imbalanced or skewed “heterogeneous”, and “cross-lingual”. The growing use of micro-text (wordplay, creative spellings, slangs) and emblematic markers (punctuations and emoticons) further increase the complexity of real-time Cyberbullying detection.

ISSN- 2394-5125 VOL 8, ISSUE 04, 2021

References

1. Rosa, Hugo, et al. "Automatic cyberbullying detection: A systematic review." *Computers in Human Behavior* 93 (2019): 333-345.
2. Cheng, Lu, et al. "Xbully: Cyberbullying detection within a multi-modal context." *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019.
3. Balakrishnan, Vimala, Shahzaib Khan, and Hamid R. Arabnia. "Improving cyberbullying detection using Twitter users' psychological features and machine learning." *Computers & Security* 90 (2020): 101710.
4. Cheng, Lu, et al. "Hierarchical attention networks for cyberbullying detection on the instagram social network." *Proceedings of the 2019 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics, 2019.
5. Salawu, Semiu, Yulan He, and Joanna Lumsden. "Approaches to automated detection of cyberbullying: A survey." *IEEE Transactions on Affective Computing* 11.1 (2017): 3-24.
6. Balakrishnan, Vimala, et al. "Cyberbullying detection on twitter using Big Five and Dark Triad features." *Personality and individual differences* 141 (2019): 252-257.
7. Singh, Vivek K., Souvick Ghosh, and Christin Jose. "Toward multimodal cyberbullying detection." Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 2017.
8. Soni, Devin, and Vivek K. Singh. "See no evil, hear no evil: Audio-visual-textual cyberbullying detection." *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018): 1-26.
9. Dadvar, Maral, and Kai Eckert. "Cyberbullying detection in social networks using deep learning based models; a reproducibility study." *arXiv preprint arXiv:1812.08046* (2018).
10. Gencoglu, Oguzhan. "Cyberbullying detection with fairness constraints." *IEEE Internet Computing* (2020).
11. Raisi, Elaheh, and Bert Huang. "Cyberbullying detection with weakly supervised machine learning." *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. 2017.
12. Dani, Harsh, Jundong Li, and Huan Liu. "Sentiment informed cyberbullying detection in social media." *Joint European conference on machine learning and knowledge discovery in databases*. Springer, Cham, 2017.
13. Dani, Harsh, Jundong Li, and Huan Liu. "Sentiment informed cyberbullying detection in social media." *Joint European conference on machine learning and knowledge discovery in databases*. Springer, Cham, 2017.
14. Cheng, Lu, et al. "Session-based Cyberbullying Detection: Problems and Challenges." *IEEE Internet Computing* (2020).
15. Kumar, Akshi, and Nitin Sachdeva. "Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis." *Multimedia Tools and Applications* 78.17 (2019): 23973-24010.
16. Haidar, Batoul, Maroun Chamoun, and Ahmed Serhrouchni. "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content." *2017 1st Cyber Security in Networking Conference (CSNet)*. IEEE, 2017.
17. Cheng, Lu, Ruocheng Guo, and Huan Liu. "Robust cyberbullying detection with causal interpretation." *Companion Proceedings of The 2019 World Wide Web Conference*. 2019.