Original Research Paper

Volume - 11 | Issue - 07 | July - 2021 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Computer Science

A STUDY ON DEEP LEARNING ALGORITHMS FOR MULTIMODAL AND MULTILINGUAL CYBERBULLYING DETECTION

Dr. Vijayakumar V* Professor& COE, Department Of Computer Science, Sri Ramakrishna College Of Arts And Science, Coimbatore-641 006, Tamilnadu, India.*Corresponding Author

Dr Hari Prasad D Professor And Head, Department Of Computer Applications, Sri Ramakrishna College Of Arts And Science, Coimbatore-641 006, Tamilnadu, India.

ABSTRACT With the increased utilization of the internet and social media platforms, can foster destructive or harmful behaviors such as cyberbullying. Cyberbullying poses significant threat to physical and mental health of the victims. There is a demand for automatic detection and prevention of cyberbullying. In Social networks, there is a big challenge to detect the cyber bullying event and to control all the cyberbullying content and languages that users post. Due to complexity of multiple languages and cross-mix languages used in cyberbullying, the detection has remained only mildly satisfying. And also recently, images and videos dominate the social feeds in addition to text messages and comments. Machine learning and deep learning techniques can be helpful to detect the bullies and can generate a model to automatically detect multi-lingual cyberbullying actions. Deep neural architectures are useful to model, learn and fuse multi-modal data for cyber bullying detection. This paper proposes a detailed review on machine and deep learning approach for detecting and preventing multimodal and multilingual cyberbullying.

KEYWORDS : Cyberbullying, Bullying detection, Multimodal, Multilingual, Cyber Safety

1. INTRODUCTION

Nowadays, Social media usage is leading people's lives and attention. With the exponential growth of social media users, particularly heavy use, is linked to depression and various other negative side effects. The users share their information with each other. Social Media Websites contain large amounts of text and/or non-text content and other information related to aggressive behavior. This leads to the growth of cyber-criminal events for example, cyberbullying which has become a worldwide epidemic. Cyberbullying has been emerged as a form of bullying by sending harmful messages that is repeated behaviour, aimed at scaring, angering or shaming those who are targeted. Examples of cyberbullying can include rumors posted on social media; embarrassing videos or pictures; and insulting, intimidating and abusive messages posted on social networks. Cyberbullying is threatening and destructive that it can lead victims to suicide attempts and cause life-long mental damage to victims. These contents need to be identified and provide the safe zone to the users.

With the increasing adverse impact of cyberbullying on society, it is necessary to find ways to detect and prevent these issues. Many researchers have lent their contributions in inventing the ways to early detect and prevent it. Most of the studies came out with text based cyberbullying detection and very few are based on image and video cyberbullying. But, Multimedia contents are the primary type of files shared on social networks (e.g., images with caption, video live streaming, and audios). Conventionally, when information from different modalities is presented together, it often reveals complementary insights about the application domain and facilitates better learning performance. Audio and video features could be used to complement textual features and improve overall detection performance. Multi-modal context aims at identifying instances of cyberbullying by leveraging multiple modalities such as textual features, spatial locations, image and visual cues, as well as the relations among sessions and also other features.

India is a multilingual country; nowadays, people use more than one languages to communicate within themselves through social media networks. Hence, it is of the utmost importance to detect cyberbullying in multiple languages. So far, research has mostly focused on English language. Bullying is not only limited to English language and occurs in other languages. Another important challenge in the multi-lingual detection is code-switching that is switching in between the languages. Detection of Cyberbullying in a multilingual world remains a challenging problem because developing language-specific sentiment lexicons is an extremely resource-intensive process. Most of the existing studies have used conventional Machine Learning (ML) models to detect cyberbullying incidents. It covers a broad range of techniques that enables systems to quickly access and learn from data, and to make decisions. It can be helpful to detect language patterns of the bullies and also it can generate a model to automatically detect cyberbullying actions. Two main challenges are identified related cyberbullying detection such as the lack of information on how multimedia can be handled in detection and building the model, and detecting the cyberbullying content in mixing of multiple languages. Recently, Deep Neural Network Based models have also been applied for the detection of cyberbullying due it improves accuracy when trained with huge amount of data. Deep Learning really shines when it comes to complex problems such as image classification, natural language processing, and speech recognition. The proposed paper presented a detailed review on multilingual cyberbullying detection approach in multi-modal data sources.

2.CYBER BULLYING DETECTION

Cyber safety means being secure online which helps to avoid the risks. There are various types of software and applications are available to help to protect against the cyber space consequences. Parental controls monitors and restricts what a person does online. It tracks the location of the device, control screen time and block explicit content and see the activity of the targeted device in real-time. There are a wide variety of programs that do such things as block and filter the content [2] such as Net Nanny (www.netnanny.com), Safe Eyes (www.internetsafety.com/safe-eyes-parental-control-software.php), CYBERsitter (www.cybersitter.com), WebWatcher (www.webwatcher.com), MMGuardian (www.mmguardian.com), Qustodio, and Kaspersky Safe Kids. Content filtering software filters objectionable, inappropriate, or illegal content. It will block to access this content and protect and provide the cyber safe online environment. It has to be deployed across all content channels such as email, web and executable software. There are many content filtering software systems available in the market including BullyBlocker, SafeChat, Facebook WatchDog, and Rethink. . It could be used by government agencies to flag patterns of cyberbullying, threatening behavior, and extremism online.

Many algorithmic techniques are utilized in the area of cyberbullying detection, mainly Machine Learning (ML), Natural Language Processing (NLP) - Lexicon-based techniques and Deep Learning techniques. Semiu Salawu presented extensive literature review, and categorized the cyber bullying detection approaches into supervised learning, lexicon-based, rule-based, and mixed-initiative approaches classes. Supervised learning-based approaches typically use classifiers such as SVM and Naïve Bayes to develop predictive models for cyberbullying detection. Lexicon-based systems utilize word lists and use the presence of words within the lists to detect cyberbullying. Rule-based approaches match text to predefined rules to identify bullying, and mixed-initiatives approaches combine human-based reasoning with one or more of the aforementioned approaches [3]. Qianjia Huang et al., discussed about the various Cyberbullying Detection Methods such as Supervised learning, Weakly-supervised learning, Lexicon based, Rule based, Mixed–initiative. Current state-of-the-art methods for cyberbullying detection combine contextual and sentiment features with text-mining approaches [4].

2.1 Lexicon Based Techniques

Lexicon based methods are based on the simple Bag-of-Words (BoW) technique. A corpus of delicate, abusive, and unpleasant words is created. The lexicon based algorithms use this corpus to check the occurrences of the words in messages to detect bullying. It counts and weighs the words that have been evaluated and tagged. The most common lexicon resources are SentiWordNet, WordNet, and ConceptNet. Lexicon based approaches are divided into the dictionary based approach and corpus-based approach, which uses statistical or semantic methods to find sentiment polarity [5]. The corpus-based approach uses a corpus and a set of cyber bullying words. Words are extracted from the corpus and compared to the set of sentiment words. It uses different statistical methods that measure semantic similarity. The dictionary-based approach as the name implies a dictionary is used by utilizing the synonym and antonym lists that are associated with dictionary words [6].

Bedoor Y. AlHarbi et al., proposed an automatic detection of cyberbullying by using sentiment analysis and lexicon approaches. After performing the data cleaning and preprocessing step, the data were classified to bullying and non-bullying [7]. Kazim Raza Talpur et al., addressed the issue of cyberbullying behaviour on Twitter platform, where users use Roman Urdu as medium of their communication. They developed supervised machine learning method and proposed a lexicon-based model with set of features derived from Twitter [8]. The texts on social media websites are written in an unstructured manner, it makes difficult for the lexicon-based approach to detect cyberbullying based only on lexicons. Lexicons are used to extract features, which are often utilized as inputs to machine learning algorithms. For example, lexicon based approaches, such as using a profane-based dictionary to detect the number of profane words in a post, are adopted as profane features to machine learning models [9].

2.2 Machine Learning Techniques

Machine Learning algorithms process the content (images, text, whatever) and identify various trends and patterns across all of those data, based on training on labeled data such as toxic or not toxic, abusive or not abusive. Machine learning techniques make automatic detection of bullying messages in social media. This could help to construct a healthy and safe social media environment.

Cyberbullying messages are detected by using machine learning approaches in the sequence of steps: data collection, feature engineering, construction of cyberbullying detection model, and evaluation of constructed cyberbullying detection models. Different supervised learning, unsupervised learning algorithms, and Reinforcement learning are used to detect cyberbullying. Supervised Learning model is built based on data which contains both set of inputs and desired outputs. Unsupervised Learning model takes set of data as input, and try to find out structure (e.g., grouping or clustering of the data). Reinforcement learning approach is concerned with taking suitable actions so as to maximize the reward in particular situation.

M. A. Al-Garadi et al., reviewed cyberbullying prediction models and identify the main issues related to the construction of cyberbullying prediction models. The data collection and feature engineering process has been elaborated, emphasis on feature selection algorithms and various machine learning algorithms for prediction of cyberbullying behaviors [10]. Gutiérrez-Esparza GO et al., presented a method to classify situations of cyber-aggression on social networks, specifically for Spanish-language users of Mexico. They applied Random Forest, Variable Importance Measures (VIMs), and OneR to support the classification of offensive comments in three particular cases of cyber-aggression such as racism, violence based on sexual orientation, and violence against women [11]. Rahat Ibn Rafq, et al., developed a cyberbullying detection system for media-based social networks, consisting of a dynamic priority scheduler, a novel incremental classifier, and an initial predictor. A multi-stage cyberbullying detection solution that drastically reduced the classification time and the time to raise alerts [12]. Van Hee C et al., proposed automatic cyberbullying detection in social media text by modelling posts written by bullies, victims, and bystanders of online bullying [13]. The author proposed a machine learning method to cyberbullying detection by making use of a linear SVM classifier exploiting a varied set of features. The approach to the annotation of fine-grained text categorized related to cyberbullying and the detection of signals of cyberbullying events.

Elaheh Raisi et al., proposed a machine learning method for simultaneously inferring user roles in harassment-based bullying and new vocabulary indicators of bullying. They also presented an automated, data-driven method for identification of harassment. The participant-vocabulary consistency model, a weakly supervised approach for simultaneously learning the roles of social media users in the harassment form of cyberbullying and the tendency of language indicators used in cyberbullying detection [14]. Love Engman described the construction of a software prototype capable of automatically identifying bullying comments on the social media platform ASKfm using Natural Language Processing and Machine Learning techniques [15]. Homa Hosseinmardi et. al. investigated the understanding and automatic detection of cyberbullying over images in media-based mobile social network, Instagram. They devised two classifiers, Naïve Bayes and linear SVM classifier separately on a sample Instagram data set consisting of manually labeled images and their associated comments [16].

V.H. Cynthia et al. detected different categories of cyberbullying e.g. blackmail, curse, defamation, insult, sexual talk, defense, harasser encouragement etc. from different sources which include the social networking site - Ask.fm, campaign donations consisting of victim's messages of cyberbullying and simulations. The data were classified using support vector machines since they work well with high-skew text classification tasks. As preprocessing steps, they applied tokenization, PoS tagging and lemmatization to the data. They carried out two classification tasks such as Cyberbullying event detection and the classification of text categories related to cyberbullying [17]. Vinita Nahar et al., proposed a method to detect cyberbullying activities on social media to identify cyberbullying messages, predators and victims. The first phase accurately detected harmful messages with semantic and weighted features. The second phase analysed social networks to identify predators and victims through their user interactions, and to present the results in a graph model [18]. Nilesh J. Uke et. al. proposed a method to segmentation and classification phases for extracting the key frames in nude images, segregation of objectionable videos, respectively. The videos were marked as porn or non-porn depending upon the judgment criteria [19]. B.Sri Nandhinia, and J.I.Sheeba, proposed a detection method to identify the presence of cyberbullying terms and classified cyberbullying activities in social network such as Flaming, Harassment, Racism and Terrorism, using Fuzzy logic and Genetic algorithm. The effectiveness of the system is increased using Fuzzy rule set to retrieve relevant data for classification from the input. Genetic algorithm is also used genetic operators like crossover and mutation for optimizing the parameters to obtain precise output [20].

2.2 Deep Learning

Deep Learning has recently emerged as an effective machine learning technique that specializes in recognizing patterns in large volumes of data. It plays a significant role in the social media analysis in noisy crises situations and to determine psychological traits based on an individual's online profile and communications.

The text-based psychological classification and demonstrates its efficacy at recognizing implicit behavioral patterns. Deep Learning learns the massive amounts of unsupervised data. It makes a valuable tool for Big Data Analytics where raw data is largely unlabeled and un-categorized [21]. Agrawal S., Awekar A. proposed deep learning based models to overcome the bottlenecks such as targeting only one particular social media platform (SMP). They investigated with machine learning models (logistic regression, support vector machine, random forest, naive Bayes) and deep neural network models (CNN, LSTM, BLSTM, BLSTM with Attention) using variety of representation methods for words (bag of character n-gram, bag of word unigram, GloVe embeddings, SSWE embeddings) [22]. N. Chandra et al., captured data sets through secured API's exposed by social network sites and stored in NoSQL databases. Tensorflow API's have been used to do predictive analysis of stored data through recursive networks. Then NLP is applied to break the text which is applied against text corpus to identify analogous data. WordNet API is being leveraged for capabilities to find synonyms [23]. Rosa at al., presented state-of-the-art in cyberbullying detection exposes that deep learning techniques. They implemented simple CNN, a hybrid CNN-LSTM and a mixed CNN-LSTM-DNN for cyberbullying detection [24].

R. Zhao and K. Mao developed semantic-enhanced marginalized denoising auto-encoder (smSDA) to tackle numerical representation learning of text messages. The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. It is able to exploit the hidden feature structure of bullying information and learn a robust and discriminative representation of text [25].

Iwendi et al., performed an empirical analysis to determine the effectiveness and performance of deep learning algorithms in detecting insults in Social Commentary. Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN) deep learning models were used. Data pre-processing steps were text cleaning, tokenization, stemming, Lemmatization, and removal of stop words. After performing data pre-processing, clean textual data is passed to deep learning algorithms for prediction [26]. Mehdi Ben et al.,used Deep Learning algorithms to extract features and patterns related to the text and concepts available in crisis related social media posts and used them to provide an overview of the crisis[27]. Bryan Perozzi et al., proposed DeepWalk, an approach for learning latent social representations of vertices using deep learning algorithms. Using local information from truncated random walks as input, this method learns a representation which encodes structural regularities [28].

Sweta Agrawal, Amit Awekar proposed the deep learning based models to overcome bottlenecks such as targeting a particular social media platform, and cyberbullying. They handled of handcrafted features of the data. They used Convolutional Neural Network (CNN), DeepNLP□ —□ Long Short Term Memory (LSTM), Bidirectional LSTM (BLSTM) algorithms with attention for cyberbullying detection [29]. Veeramallu Naga Srinivas and, Veerendra Bethimeedi, investigated stacked de-noising auto-encoder (SDA) deep learning method. They developed a new text representation model based on a variant of SDA called marginalized stacked de-noising auto-encoders (mSDA), which adopts linear instead of nonlinear projection to accelerate training and marginalizes infinite noise distribution in order to learn more robust representations. The Semantic-enhanced Marginalized Stacked Denoising Autoencoder is able to learn robust features from BoW representation in an efficient and effective way. These robust features are learned by reconstructing original input from corrupted (i.e., missing) ones. The new feature space can improve the performance of cyberbullying detection even with a small labeled training corpus [30]. M. Ptaszynski, et al, proposed a novel method to automatic cyberbullying detection based on Convolutional Neural Networks and increased Feature Density [31].

3. MULTIMODAL DEEP LEARNING ALGORITHMS FOR CYBERBULLYING DETECTION

Deep networks have been successfully applied to unsupervised feature learning for single modalities (e.g., text, images or audio) to detect the cyber bullying activities. Cyberbullying instances are not only taking place using texts, but also image, audio and video features and other features plays an important role in spreading cyberbullying.

The Instagram media-based social network is well-suited to cyberbullying prediction since there is an initial posting of an image typically with an associated text caption, followed later by the text comments that form the basis of a specific cyberbullying incident. Several important multimodal features are extracted from the initial posting data for automated cyberbullying prediction, including profanity and linguistic content of the text caption, image content, as well as social graph parameters and temporal content behavior [32]. Lu Cheng et al., presented a detailed review on cyberbullying detection within a multi-modal context by exploiting social media data in a collaborative way. The XBully cyberbullying detection framework reformulates multi-modal social media data as a heterogeneous network and then aims to learn node embedding representations upon it. It identifies representative mode hotspots to handle diverse feature types and then jointly maps both attributed and nominal nodes in a heterogeneous network into the same latent space by exploiting the cross-modal correlations and structural dependencies [1]. Machine learning techniques can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyberbullying multimedia content. Devin Soni and Vivek Singh built an automatic cyberbullying classifier using machine learning with the text, audio & visual, and text + audio + visual features [33].

Vivek K. Singh, et al., proposed the use automated image and text analysis APIs for multimodal (visual and textual) cyberbullying detection. They designed predictive models for automatic cyberbullying detection, which include the use of visual features (e.g. gender, race, nudity, portrayed emotions etc.) to complement textual features. They closely follow recent works using textual features for cyberbullying detection as well as utilizing multiple channels of data (e.g. social and textual) for cyberbullying detection. They built an automatic machine learning based classifier to detect media sessions containing cyberbullying [34]. Naveen Kandlapalli et al., used a combination of techniques like image analysis and text analysis used in order to curb the Cyberbullying attacks and categorized the text and image input of the user as abusive or non-abusive[35]. Krishna B. Kansara and Narendra M. Shekokar proposed a framework had two main modules such as Abusive image detection and Abusive text detection deployed for the detecting negative online interactions in terms of abusive contents carried out through text messages as well as images [36]. Shardul Suryawanshi, et al., created the MultiOFF multimodal meme dataset for offensive content detection dataset. They subsequently developed a multimodal classifier. They used an early fusion technique to combine the image and text modality and compare it with a text- and an image-only baseline to investigate its effectiveness [37]. L Cheng, K Shu et al., Proposed a model consists of two main components such as representation learning network that encodes the social media session by exploiting multi-modal features, e.g., text, network, time and multi-task learning network that simultaneously fits the comment inter-arrival times and estimates the bullying likelihood based on a Gaussian Mixture Model. The model jointly optimizes the parameters of both components to overcome the shortcomings of decoupled training [38].

K. Wang et al., proposed a multi-modal detection framework that takes into multi-modal information (e.g., image, video, comments, time) on social networks. They used three modules to extract modality features in a social network and fused multiple data types. The first module used bidirectional LSTM with attention to extracting the post's characteristics. They introduced hierarchical attention networks to apply at word and comment level. Then they used MLP to encode other meta information, such as video and image. By processing different modal data, they constructed a multi-modal cyberbullying detection framework and utilize the framework [39]. Devin Soni focused on audio-visual-textual cyberbullying detection. They identified textual, audio, and visual features relevant for cyberbullying detection [33]. Areej Al-Hassan and Hmood Al-Dossari presented a comprehensive study on the usage of text mining in social networks. They investigated some challenges which can be a guide for the implementation of Arabic hate speech detection model [41].

3.1 Multimodal Features Extraction and Information Fusion

Multiple data sources are semantically correlated and provide complementary information to each other. The patterns aren't visible when working with individual modalities on their own but the fusion of heterogeneous data produce more robust predictions. The task can mainly be divided into two steps such as individual feature learning, information fusion. So, multiple modalities data in social media such as text, image, video, audio, and speech features are learned to prediction. Feature extraction from one source is independent from another. The features are extracted from individual sources of information by building models that best suit the type of data. For example, the features extracted from images texts (super imposed text or image text) are in the form of edges. Token features extracted from text data, latent topics from captions and visual cues features extracted from images and videos (image-specific features such as color histogram). After all the features important for prediction are extracted from all data sources.

Many feature selection methods rely on machine learning classifiers which may not be robust across datasets. Then combine the different features into one shared representation. Deep neural networks have been successfully applied to unsupervised feature learning for single modalities—eg. text, images or audio. These fusion improves social media network's predictive ability.

4. MULTILINGUAL CYBERBULLYING DETECTION

Cyberbullying is not location and language specific, i.e., it occurs worldwide and across different languages. A lot of research work proposed solutions for detecting cyberbullying in English language and a few more languages, but very less works covered cyberbullying in multilingual language detection.

Braja Gopal Patra et al., presented overview of the shared task on sentiment analysis of code-mixed data pairs of Hindi-English and Bengali-English collected from the different social media platform. They used word and character level n-grams as features and SVM for sentiment classification [42]. Tarwani S., et al., created the Hinglish Cyberbullying Comments (HCC) labeled dataset consisting of comments from social media networks such as Instagram and YouTube. They also developed eight different machine learning models for sentiment classification in-order to automatically detect incidents of cyberbullying [43]. Anisha Datta et al., provided datasets in three languages – English, Hindi and Bengali. The task was to classify each instance of the test sets into three categories such as Overtly Aggressive (OAG), Covertly Aggressive(CAG) and Non-Aggressive (NAG). They used three different models for three different languages. They used Tf-Idf vectorizer to vectorize the word-tokens. For English dataset, they used the XGBoost classifier followed by the bagging method. For Hindi dataset, they used the Gradient Boosting classifier and many different types of features like aggressive words lexicon, sentiment scores, parts of speech tags etc. They used the Gradient Boosting Classifier for Bengali dataset [44].

Sandip Modha et al., identifed the level of aggression from the User-Generated contents within Social media written in English, Devnagiri Hindi and Romanized Hindi. Aggression levels are categorized into three predefined classes namely: 'Overtly Aggressive', 'Covertly Aggressive'and 'Non-aggressive'. They developed a multi-class classifier which classifies User-generated content into these pre-defined classes. They experimented with standard machine learning classifiers and deep learning models for the multi-class classification problem. They have also found that hyper-parameters of the deep neural network are the keys to improve the results [45]. Igor Mozetic et al., analyzed a large set of manually labeled tweets in different languages, use them as training data, and construct automated classification models. They analyzed a set of over 1.6 million Twitter posts, in 13 European languages, labeled for sentiment by human annotators. The labeled tweets are used as training data to train sentiment classifiers for different languages [46]. Batoul Haidar et al., proposed a solution for the problem of cyberbullying in both English and Arabic languages. This system employed machine learning and two toolkits were tested for ML, Dataiku DSS and WEKA. The decision was to use WEKA toolkit because it supports Arabic language [47].

Pawar, Rohit, S. MS proposed a Multilingual Cyberbullying Detection System for the detection of cyberbullying in multiple languages (English, Hindi, and Marathi). They used two techniques, namely, Machine Learning-based such as Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Stochastics Gradient Descent (SGD), and Logistic Regression (LR), and Lexicon-based, to classify the input data as bullying or non-bullying [48]. Kumar, A., and Sachdeva, N. focused on cyberbullying detection in the code-mix data, specifically the Hinglish, which refers to the juxtaposition of words from the Hindi and English languages. They explored the problem of cyberbullying prediction and proposed MIIL-DNN, a multi-input integrative learning model based on deep neural networks. MIIL-DNN combines information from three sub-networks to detect and classify bully content in real-time code-mix data. It takes three inputs, namely English language features, Hindi language features (transliterated Hindi converted to the Hindi language) and typographic features, which are learned separately using sub-networks (capsule network for English, bi-LSTM for Hindi and MLP for typographic). These are then combined into one unified representation to be used as the input for a final regression output with linear activation. The advantage of using this model-level multi-lingual fusion is that it operates with the unique distribution of each input type without increasing the dimensionality of the input space. The robustness is validated on two datasets created by scraping data from the popular social networking sites, namely Twitter and Facebook[40]. Kumar, A., and Sachdeva, N. presented the use and application of soft computing techniques for cyberbullying detection on social multimedia utilizing a meta-analytic approach in order to integrate, interpret and critically analyze the findings in the original studies for expounding novel approaches to achieve comparable and effectual results pertaining to the defined research domain [50]. Thiago Galery and Efstathios Charitos aligned Hindi and English vectors using pre-computed SVD matrices that pulls representations from different languages into a common space. They used a classification pipeline with two conditions such as single language condition regarded foreign (Hindi) words as Out of Vocabulary (OOV) tokens and multi-language condition aligned English and Hindi fastText vectors via SVD matrices [51].

5. MULTIMODAL-MULTILINGUAL DEEP LEARNING FRAMEWORK

This proposed Multimodal Multilingual cyberbullying detection framework consists of three key phases, such as Data collection, Preprocessing and Detection of cyberbullying is shown in the Fig.1.

The first phase collects the input data from the social media network websites or from the benchmark dataset libraries. In the Pre-Processing phase, noises are removed to improve the quality of the subsequent analytical steps. In the text data processing like removes stop words, unwanted characters, etc.. The features are extracted from the multi-lingual data source. Then image and video features are also extracted. Finally the all the data are fused and applied to the classifier to improve the prediction accuracy. In the Detection and Classification, deep learning classifier is applied to detect patterns used by bullies and their victims, and develop rules to automatically detect cyberbullying content.

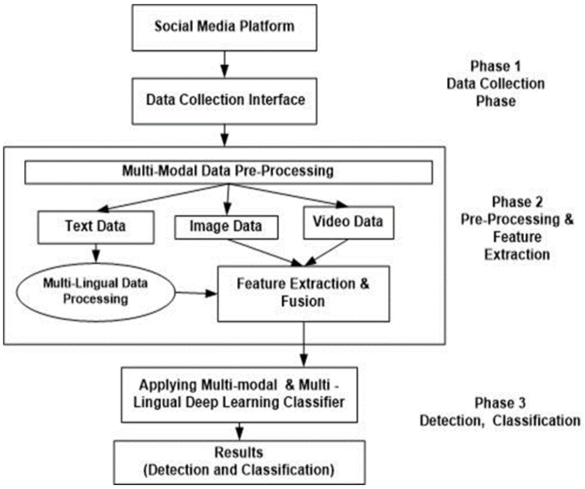

Figure 1. Multimodal-Multilingual Deep Learning Framework

6.CYBERBULLYING DATASETS

The process of detecting cyberbully activities begins with input dataset from social network. There is a vast amount of rich social network data available on the Web. The pre-existing secondary data collected from popular benchmark websites. All these datasets are manually labelled and publicly available. It covers all types of hetro-genious data (text, image, video) for testing and analysis. The Formspring.me, MySpace data, Ask.fm, Instagram, twitter and Vine data have been collected from the popular benchmark datasets repository provided by researcher. The repositories are:
https://sites.google.com/site/cucybersafety/home/cyberbullying-detection-project/dataset
http://research.cs.wisc.edu/bullying/data.html
http://www.chatcoder.com/DataDownload
https://github.com/Mrezvan94/Harassment-Corpus
https://github.com/ENCASEH2020/hatespeech-twitter
The real time data will be collected from social networking sites with Python API. Instagram provides an official application program interface (API), which makes it convenient for us to download the data needed. Twitter API is used to get the tweet text such as https://api.twitter.com/1/statuses/show/850660404770590720.json.

7. RESEARCH ISSUES & CHALLENGES

Existing work on cyberbullying detection is mainly based on uni-modal such as text, image, and uni-language like English, Hindi. But, multimodal and multilingual cyberbullying detection also encounters several challenges:

- Data Gathering and Labelling: To train the machine learning and deep learning model, data are collected from different sources. And also the multi-languages have limited resources publically available.
- Data Pre-processing: Since, data obtained from multiple sources; the feature selection and extraction of the each resource also a complex problem.
- Data fusion: Multi-feature data fusion and processing is challenge task.
- Language Training: Dynamic algorithms are required to detect new slang and abbreviations related to cyberbullying behavior and keep updating the training processes of machine/deep learning algorithms by using newly introduced words in Multilanguage.
- Algorithms: Deep learning architectures remain unexplored in multimodal and multi-lingual cyberbullying detection in social media websites.
- Storage: Lack the capability to handle cyberbullying big data.

8. CONCLUSIONS

Early detection of cyberbullying content becomes of utmost importance due to growing number of incidents in online social media. The precise type of cyberbullying activity detection helps to social media users, parents, government or other social welfare organization to identify the cyberbullying events well in advance and take necessary actions to prevent the users of the social network from becoming victims. Cyberbullying instances are not only taking place using texts, but also audio and video features play an important role in spreading cyberbullying. The multimodal and multi-lingual approach can play a significant role in reducing cyberbullying in cross language and improving the quality of life of individuals who are affected by cyberbullying each year. This paper has provided a multilingual cyberbullying detection in multi-modal data sources. The paper also summarized the data repository used for cyberbullying detection. Finally, the issues and research challenges were described and discussed.

ACKNOWLEDGMENT

This research work was funded by the ICSSR under IMPRESS Scheme. The author also thanks the ICSSR and Sri Ramakrishna College of Arts and Science who provided the support for doing the research.

REFERENCES

[1] Lu Cheng, Jundong Li , Yasin N. Silva , Deborah L. Hall. (2019) XBully: Cyberbullying Detection within a Multi-Modal Context. In The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), Melbourne, VIC, Australia. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291037

[2] John Sammons, Michael Cross, (2017), Chapter 10 - Protecting your kids, Editor(s): John Sammons, Michael Cross, The Basics of Cyber Safety, Syngress, Pages 201-227, ISBN 9780124166509. (http://www.sciencedirect.com /science/article/pii /B9780124166509000103)

[3] Salawu, Semiu & He, Yulan & Lumsden, Joanna. (2017). Approaches to Automated Detection of Cyberbullying: A Survey. IEEE Transactions on Affective Computing. PP. 1-1. 10.1109/TAFFC.2017.2761757.

[4] Qianjia Huang, Jianhong Zhang, Diana Inkpen, David Van Bruwaene, (2018), Cyberbullying Intervention Interface Based on Convolutional Neural Networks, Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pages 42–51 Santa Fe, USA.

[5] Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan,Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba,Haruna Chiroma, Hasan Ali Khattak , And Abdullah Gan, (2019), Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges, in IEEE Access, vol. 7, pp. 70701-70718, doi: 10.1109/ACCESS.2019. 2918354.

[6] Bedoor Y. AlHarbi , Mashael S. AlHarbi , Nouf J. AlZahrani , Meshaiel M. Alsheail , Jowharah F. Alshobaili and Dina M. Ibrahim, (2019), Automatic Cyber Bullying Detection in Arabic Social Media, International Journal of Engineering Research and Technology. ISSN 0974-3154, 12(12), pp. 2330-2335.

[7] Kumar, A., Sachdeva, N. (2019), Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. Multimed Tools Appl 78, 23973–24010 https://doi.org/10.1007/s11042-019-7234-z,

[8] Kazim Raza Talpur, Siti Sophiayati Yuhaniz, Nilam Nur binti Amir Sjarif, Bandeh Ali, (2020), Cyberbullying Detection in Roman Urdu Language Using Lexicon Based Approach, International Journal of Advance Science and Technology, 29 (10S), pp. 786-800.

[9] K. Reynolds, A. Kontostathis, and L. Edwards, (2011), Using machine learning to detect cyberbullying," in Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA), pp. 241–244.

[10] Thiago Galery, Efstathios Charitos, (2018), Aggression Identification and Multi-Lingual Word Embeddings, Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pages 74–79, Santa Fe, USA.

[11] Gutiérrez-Esparza GO, Vallejo-Allende M, Hernández-Torruco J. (2019), Classification of Cyber-Aggression Cases Applying Machine Learning. Applied Sciences.; 9(9):1828

[12] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. (2018), Scalable and Timely Detection of Cyberbullying in Online Social Networks. In SAC 2018: SAC 2018: Symposium on Applied Computing , Pau, France. ACM, New York, NY, USA, 10 pages. https://doi.org/10. 1145/3167132.3167317

[13] Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, et al. (2018), Automatic detection of cyberbullying in social media text. PLoS ONE 13(10): e0203794. https://doi.org/10.1371/journal.pone.0203794

[14] Elaheh Raisi, Bert Huang, (2017), Cyberbullying Detection with Weakly Supervised Machine Learning, ASONAM'17, Sydney, Australia, ACM, http://dx.doi.org/10.1145 /3110025.3110049

[15] Love Engman, (2016), Automatic Detection of Cyberbullying on Social Media, Master's Thesis in Computing Science, Umea University, SWEDEN

[16] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra, (2015), Detection of Cyberbullying Incidents on the Instagram Social Network, Association for the Advancement of Artificial Intelligence, ARXIV.

[17] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes,, Bart Desmet, Guy De Pauw, Walter Daelemans and Veronique Hoste, (2015), Detection and Fine-Grained Classification of Cyberbullying Events, Proceedings of Recent Advances in Natural Language Processing, pages 672–680, Hissar, Bulgaria.

[18] Vinita Nahar, Xue Li, Chaoyi Pang, (2013), An Effective Approach for Cyberbullying Detection, Communications in Information Science and Management Engineering, 3 (5), PP. 238-247

[19] Nilesh J.Uke, Dr. Ravindra C. Thool, (2012), Detecting Pornography on Web to Prevent Child Abuse – A Computer Vision Approach, International Journal of Scientific & Engineering Research, pp. 1-3.

[20] B.Sri Nandhinia, J.I.Sheeba, (2015) Online Social Network Bullying Detection Using Intelligence Techniques, Procedia Computer Science, 45, 485 – 492, International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

[21] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald and Edin Muharemagic, (2015) , Deep learning applications and challenges in big data analytics, Journal of Big Data 2:1, 1-21.

[22] Agrawal S., Awekar A. (2018) Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. In: Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (eds) Advances in Information Retrieval. ECIR 2018. Lecture Notes in Computer Science, vol 10772. Springer, Cham. https://doi.org/10.1007/978-3-319-76941-7_11

[23] N. Chandra, S. K. Khatri and S. Som, (2018) Cyberbullying Detection using Recursive Neural Network through Offline Repository, 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2018, pp. 748-754, doi: 10.1109/ICRITO.2018.8748570.

[24] H. Rosa, D. Matos, R. Ribeiro, L. Coheur and J. P. Carvalho, (2018), A "Deeper" Look at Detecting Cyberbullying in Social Networks," International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489211.

[25] R. Zhao and K. Mao, (2017), Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder, in IEEE Transactions on Affective Computing, vol. 8, no. 3, pp. 328-339, doi: 10.1109/TAFFC.2016.2531682.

[26] Iwendi, C., Srivastava, G., Khan, S. et al. (2020), Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems, https://doi.org/10.1007/ s00530-020-00701-5

[27] Mehdi Ben Lazreg Morten Goodwin Ole-Christoffer Granmo, (2016), Deep Learning for Social Media Analysis in Crises Situations, The 29th Annual Workshop of the Swedish Artificial Intelligence Society (SAIS),Malmö, Sweden

[28] Bryan Perozzi , Rami Al-Rfou and Steven Skiena, (2014) DeepWalk: online learning of social representations, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 701-710.

[29] Sweta Agrawal, Amit Awekar, (2018), Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms, arXiv:1801.06482v1 [cs.IR].

[30] Veeramallu Naga Srinivas, Veerendra Bethimeedi, (2017), Detection of Text based Cyberbullying using Semantic Enhanced Marginalized Denoising Autoencoder Learning, International Journal of Computer Science and Mobile Computing, 6(8), pg.89 – 94.

[31] M. Ptaszynski, JKK. Eronen, F. Masui, (2017) Learning Deep on Cyberbullying is Always Better than Brute Force, in: Proceedings of the Linguistic and Cognitive Approaches to Dialog Agents (LaCATODA 2017), CEUR Workshop Proceedings, vol. 1926, pp. 3-10.

[32] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra, (2016), Detection of Cyberbullying Incidents on the Instagram Social Network, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[33] Devin Soni and Vivek Singh. (2018). See No Evil, Hear No Evil: Audio-Visual-Textual Cyberbullying Detection. Proceedings of the ACM on Human-Computer Interaction 2, CSCW, Article 164 (November 2018), 25 pages. https://doi.org/10.1145/3274433

[34] Vivek K. Singh, Souvick Ghosh, Christin Jose, (2017), Toward Multimodal Cyberbullying Detection, In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17). Association for Computing Machinery New York, NY, USA, 2090–2099. DOI:https://doi.org/10.1145/ 3027063.3053169

[35] Naveen Kandlapalli, Shobha Shinde, Priyanka Shriramoji, Pooja Uke, Supriya Chaudhary, (2017), Defending Mechanism for Cyber Bullying, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2(2), pp. 716-719

[36] Krishna B. Kansara and Narendra M. Shekokar, (Feb 2015), A Framework for Cyberbullying Detection in Social Network,", International Journal of Current Engineering and Technology, 5(1), pp. 494-498

[37] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan and Paul Buitelaar . (2020), Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text, Proceedings of the LREC 2020 Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2,).

[38] L Cheng, K Shu, S Wu, YN Silva, DL Hall, H Liu, (2020) Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model, The 29th ACM International Conference on Information and Knowledge Management.

[39] K. Wang, Q. Xiong, C. Wu, M. Gao and Y. Yu, (2020), Multi-modal cyberbullying detection on social networks, International Joint Conference on Neural Networks (IJCNN), Glasgow, United Kingdom, pp. 1-8, doi: 10.1109/IJCNN48605. 2020.9206663.

[40] Kumar, A., Sachdeva, N. (2020). Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. Multimedia Systems https://doi.org/10.1007/s00530-020-00672-7.

[41] Areej Al-Hassan and Hmood Al-Dossari (2019), Detection Of Hate Speech In Social Networks: A Survey On Multilingual Corpus, Dhinaharan Nagamalai et al. (Eds) : COSIT, AIAPP, DMA, SEC, pp. 83–100.

[42] Braja Gopal Patra, Dipankar Das, and Amitava Das, (2017), Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL Code-Mixed Shared Task @ICON, arXiv:1803.06745.

[43] Tarwani S., Jethanandani M., Kant V. (2019) Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification. In: Singh M., Gupta P., Tyagi V., Flusser J., Ören T., Kashyap R. (eds) Advances in Computing and Data Sciences. ICACDS. Communications in Computer and Information Science, vol 1046. Springer, Singapore. https://doi.org/10.1007/978-981-13-9942-8_51

[44] Anisha Datta , Shukrity Si , Urbi Chakraborty , Sudip kumar Naskar, (2020) Spyder: Aggression Detection on Multilingual Tweets, Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 87–92, Language Resources and Evaluation Conference (LREC 2020), Marseille.

[45] Sandip Modha, Prasenjit Majumder, Thomas Mandl, (2018), Filtering Aggression from the Multilingual Social Media Feed, Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, New Mexico, USA

[46] Igor Mozetic, Miha Grcar, and Jasmina Smailovic.(2016), Multilingual twitter sentiment classification: The role of human annotators. PloS one, 11(5):e0155036.

[47] Batoul Haidar, Maroun Chamoun, Ahmed Serhrouchni, (2017), A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning, dvances in Science, Technology and Engineering Systems Journal, 2(6), 275-284

[48] Rohit S. Pawar, (2019), Multilingual Cyberbullying Detection System", A thesis, Department of Computer Science Indianapolis, Indiana May