Cyberbullying and Cyberviolence Detection: A Triangular User-Activity-Content View

Shuwen Wang, Xingquan Zhu*, *Senior Member, IEEE,* Weiping Ding, *Senior Member, IEEE,* and Amir Alipour Yengejeh

***Abstract*—Recent years have witnessed the increasing popularity of mobile and networking devices, as well as social networking sites, where users engage in a variety of activities in the cyberspace on a daily and real-time basis. While such systems provide tremendous convenience and enjoyment for users, malicious usages, such as bullying, cruelty, extremism, and toxicity behaviors, also grow noticeably, and impose significant threats to individuals and communities. In this paper, we review computational approaches for cyberbullying and cyberviolence detection, in order to understand two major factors: (1) what are the defining features of online bullying users, and (2) how to detect cyberbullying and cyberviolence. To achieve the goal, we propose a User-Activities-Content (UAC) triangular view, which defines that users in the cyberspace are centered around the UAC triangle to carry out activities and generate content. Accordingly, we categorize cyberbully features into three main categories: (1) user centered features, (2) content centered features, and (3) activity centered features. After that, we review methods for cyberbully detection, by taking supervised, unsupervised, transfer learning, and deep learning *etc.*, into consideration. The UAC centered view provides a coherent and complete summary about features and characteristics of online users (their activities), approaches to detect bullying users (and malicious content), and helps defend cyberspace from bullying and toxicity.**

***Index Terms*—Cyberbullying, social network, natural language processing, classification, clustering**

I. Introduction

Bullying is an aggressive and intentional act or behavior frequently conducted by one or a group of individuals against a victim, who often cannot defend him/herself [1]. Bullying can be carried out in many types of forms such as verbal bullying, physical attack, sexual humiliate, social isolate, psychological torment [2]. Cyberbullying is another evolution method from direct physical bullying to electronic devices, *i.e.* in a cyberspace. National Crime Prevention Council defines cyberbullying as "similar to other types of bullying, except it takes place online and through text messages sent to cell phones. Cyberbullies can be classmates, online acquaintances, and even anonymous users, but most often they do know their victims [3]". Similar definitions have also been found in several other related online incidents, such as cyberstalking, cyberharassment, online cruelty, or online cruelty in general [4], [5]. In many occasions, online cruelty is also referred to as cyberbullying [4], however, some research considers cyberbullying being part of online cruelty activities. Other online forms of digital activities such as online harassment and online sexual harassment should also be considered as online cruelty [5].

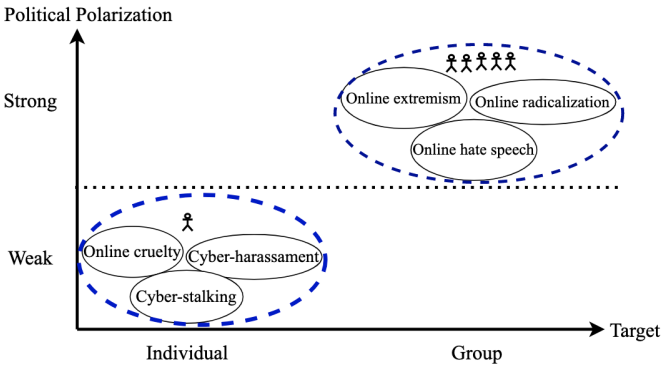


Fig. 1: A simple comparison between Cyberbullying and Cyberviolence *w.r.t.* targeted audience and political polarization.

A. Cyberbullying vs. Cyberviolence

Cyberviolence is another term relevant but different from cyberbullying. A key difference between them is that the former targets on a group of individuals with strong political preference, whereas the latter is more focused on individuals, as shown in Fig. 1. A summary and categorization between cyberbullying and cyberviolence are also reported in Table I.

In this paper, we consider online cruelty activities along with cyberharassment and cyberstalking part of cyberbullying. Cyberharassment, in general, refers to online interpersonal attacks which occur repetitively, intrusively and provokes anxiety [6]. Cyberstalking is to stalk or harass an individual, group or organization with electronic equipment [7], [8].

Online extremism, online hate speech and online radicalization are forms of cyberviolence. Online extremism is an express of extreme views of hatred toward some group using internet technologies to "advocate violence against, separation from, defamation of, deception about or hostility towards others" [9]–[11]. Online hate speech is an expression of conflict between different groups within and between societies based on race, religion, ethnicity, sexual orientation, disability,

S. Wang, X. Zhu, and A. Yengejeh are with the Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA. E-mail: {swang2020, xzhu3, aalipouryeng2018}@fau.edu. *Contacting author: Xingquan Zhu.

W. Ding is with the School of Information Science and Technology, Nantong University, Nantong, China. E-mail: ding.wp@ntu.edu.cn.

Manuscript received December 22, 2021; revised January 2022, April 2022.

TABLE I: Summary of cyberbullying and cyberviolence

Category	Subcategory	Initiator	Target	Politics	Interaction	Objective
Cyberbullying	Online cruelty Cyber-harassment Cyber-stalking	Individual	Individual	Weak	Frequent	Mental and emotion damage
Cyberviolence	Online extremism Online hate speech Online radicalization	Group	Group	Strong	Rare	Broadcast ideas; Alteration of belief

or gender. Radicalization is a process in which individuals or groups oppose the political, social, or religious status with increasingly radical views. There exists a strong dependency between cyberbullying and digital media such as hurtful images and comments where those contents can remain online accessible to public before they are reported and deleted. The explosive development of Internet technologies provides people more and more opportunities to be exposed to online information, from daily life to social interaction.

Although a large number of audience and users are attracted by continuous and immediate online social media which prompts the wide, quick spread of online contents, due to identification difficulty and loose supervision of the overall internet environment, cyberbullying has become unscrupulous among which, some even mislead to cybercrime, and hate speech. One of the most distinguishing features of cyberbullying is that victims can hardly find an effective solution to get away from it. In other words, even a one time bullying action can lead to continuous ridicule and humiliation for victims, which is possible to result in feelings of powerlessness for the victims. Besides, because of the anonymity feature of cyberbullying, failing to recognize the identity of bully increases feeling of frustration and powerlessness of the victims. According to National Center for Education Statistics, among students ages 12-18 who reported being bullied at school during the school year, 15% were bullied online or by text [12]. Victims of cyberbullying generally develop certain psychological problems such as anxiety, depression, poor performance or committing suicide. Therefore, early cyberbullying detection becomes the utmost important.

B. Computational Approaches for Cyberbullying Detection

Computational approaches, such as machine learning, have been used for automatic detection of cyberbullying, and a general framework used by these methods is summarized in Fig.2. Online social networks contain useful information from user posts, interactions, *etc*. Those information can be extracted as features which will then be fed into models to learn and make prediction. One typical way of currently existing studies is using machine learning classifier, combined with text mining, as supervised learning approach in social medias [13], [14]. For example, Random Forest is used based on user personality features decided by Big Five and Dark Triad models for Twitter cyberbullying detection and achieves up to 96% precision [15]. Deep learning models Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Recurrent

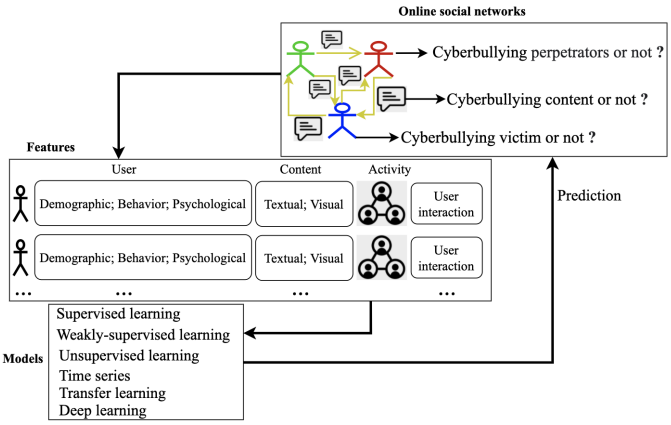


Fig. 2: A conceptual view of computational framework for Cyberviolence and Cyberbullying detection

Neural Network (RNN) are applied in detecting insults in Social Commentary and their results indicates that deep learning models is more effective when compared with other traditional methods [16].

Many methods exist for cyberbullying detection. While technical solutions vary, these approaches often face and address similar challenges. First of all, due to the limitations of social medias, adequate and informative data collection cannot be guaranteed and users can modify, delete their social media content at any time, which means data availability changes with time. Traditional data cleansing methods may mistakenly filter out important information, which will hurt the ability of machine learning models to discriminating bullying and normal expressions. Besides, unstructured content can be found in multiple language formats and styles, which is often grammatically inaccurate. Another common challenge is that cyberbullying content is quite rare and only limited amount of cyberbullying will be found even in large datasets, which poses a great challenge for traditional machine learning models.

C. Existing survey, difference, contributions

A handful of survey already exist to focus on cyberbullying detection from different perspectives. Techniques utilized in the field are briefly summarized as two categories, machine learning methods and Natural Language Processing (NLP) methods [17]. A survey of the main algorithms for text mining with a focus on cyberbullying detection is proposed in which the most common methods such as vector space modelling is discussed [18]. Supervised learning, lexicon-based, rule-based, and mixed-initiative classes are proposed in order to

categorize approaches used for cyberbullying detection and features in those reviewed paper are separated into four main groups, content-, sentiment-, user- and network-based features [19]. Challenges ranging from definition of cyberbullying, collecting data, feature selection and model selection are surveyed and suggestions for tackling those limitations are proposed in a survey on automated cyberbullying detection [20].

Although cyberbullying research has received increasing attention for more than a decade, majority of current survey papers focus on introducing methods used and certain general feature categorization, which is limited to provide detailed, precisely instructions for researchers and practitioners in the field. Many key questions remain unanswered. For example, what are the source of information for cyberbullying detection? what are the essential components cyberbullying detection, and how these components interplay with each other? Finally, what are available approaches for cyberbullying detection.

Motivated by the above, in this paper, we propose a triangular user-activity-content view of cyberbullying detection. The unique view defines that users in the cyberspace are centered around this triangle to carry out activities and generate content. The interplay between these three factors (users, activities, and content) essentially sheds the light for designing cyberbullying detection algorithms. Following this triangular view, we further review methods for cyberbully detection, by taking supervised, unsupervised, transfer learning, and deep learning, etc. into consideration. In addition to the introduction of these approaches, we also explain how they focus on different types of features and the common combinations of different types of features, with respect to the proposed UAC triangular view.

D. Literature Review Process

To ensure comprehensiveness and completeness, we carried out a literature search strategy to collect papers and set up a selection criteria to choose final reference for this report. E-databases of IEEE Xplore, Science Direct, PubMed, Google scholar and arXiv were explored in order to collect adequate literature. Snowball sampling method was conducted from the reference of chosen studies to find more related papers. At first, we used keywords like Cyberviolence detection, Cyberbullying, Cyberbullying detection, Cyberbullying prediction to search records. The original research did not bring us enough papers, therefore, we expanded research with more keywords, like Online cruelty, Cyber-harassment, Cyber-stalking, Online extremism, Online hate speech and Online radicalization. Papers containing the above mentioned keywords in tile, abstract, keywords and published in English language are included as our final paper resource. Studies from short paper summaries, magazines, non-machine learning methods included, incomplete studies and not published in English language are excluded in this research. The complete selection process of reviewed articles in this research as shown in Fig. 3. Table II reports the number of papers selected in the review study.

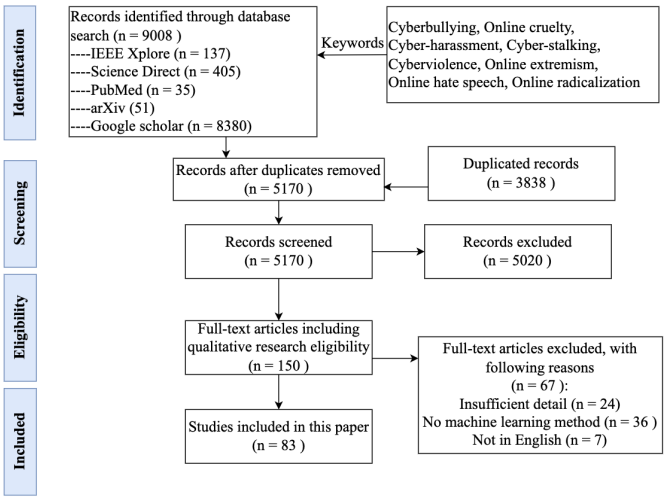


Fig. 3: The systematic literature review process of the proposed study

TABLE II: Summary of final paper numbers for review study

Feature	# of paper	Detection method	# of paper
		Supervised learning	27
User	15	Weakly-supervised learning	3
Activity	13	Unsupervised learning	2
Content	28	Time series	7
		Transfer learning	2
		Deep learning	16

II. User-Activity-Content Triangle

In cyberspace, the impact of cyberbullying and other harmful materials/activities are facilitated through three major parties: User, Activity and Content, which form a UAC triangle. Fig. 4 shows the relationship between User, Activity and Content. A user is the main part of cyberbullying who carries out activities and form interactions with him/herself and other users. The content produced from user activities is the consequence causing people to suffer from cyberbullying and affecting subsequent user activities. Despite of the actual forms of online harmful activities, *e.g.* texts, images, symbols, slang *etc.*, it is always around this UAC triangle.

Using UAC triangular view, we can further understand the difference between cyberbullying and cybervoilence, as summarized in Table I. Cyberviolence is similar to cyberbullying, but unlike cyberbullying the majority of which happen between individuals (both the activity initiator and target). In most scenarios, online extremist, online hate speech and online radicalization are started by a group of people and aim the abuse at a collective identity for example, specific groups. Once the activity is initiated, the interaction between activity conductor and victims in cyberbullying is more frequent than that in cyberviolence. In general, cyberbullying conductors is aware of their behavior as well as the possible results of them, however, cyberviolence materials are usually considered as educational rather than offensive [11], [21]. In addition, as indicated in our table, cyberviolence activities usually

have stronger political overtones as shown in Fig 1. than cyberbullying because the objective of them is trying to spread their beliefs, ideas and alter others' belief. Most cyberviolence activities do not directly promote violence, and exposure does not necessarily cause trauma or other adverse effects [22]. On the contrary, both mental and physical damage can be observed immediately from cyberbullying victims.

Based on the UAC triangle, we propose a cyberbullying feature taxonomy in Fig. 5, where features commonly used in cyberbullying are categorized into three groups, User feature, Activity feature, and Content feature.

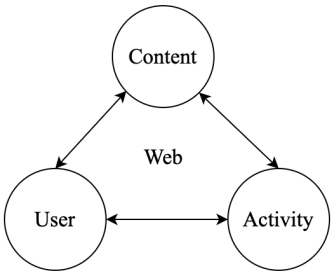


Fig. 4: The proposed UAC triangle to characterize relationships between users, user activities, and content generated by users.

III. User-centered Features

User centered features include three main aspects, demographics, behaviors, and physiologic, in an increasing level of intelligence. In an ideal case, such features can help characterize users by answering: who are the authors (gender, locations, *etc.*), what are their behaviors (online active/inactive, *etc.*), and what are their personality traits (extroversion/openness, *etc.*). Table III summarize main user-centered features, including sub-features, strength and weakness of these features.

A. Demographic

Demographic features or user characteristics can be defined as one's personal information such as gender, age, race, education level and profession on the social platforms which usually can be found from user account profile. For online social networks, user profile is a collection of information associated with a user which includes key information used to identify individuals, such as name, portrait photos, number of followers. User profiles most often appear on social media sites such as Facebook, Instagram, and LinkedIn; they serve as individuals' voluntary digital identities, highlighting their main characteristics and characteristics. Such personal profile information is used as the basis for grouping users, sharing content, and recommending or introducing people who is engaged with direct interaction [24]. However, their reliability has caused major concern because social networks allow individuals to create unlimited personal profiles, making it easy to create false or inconsistent personal information.

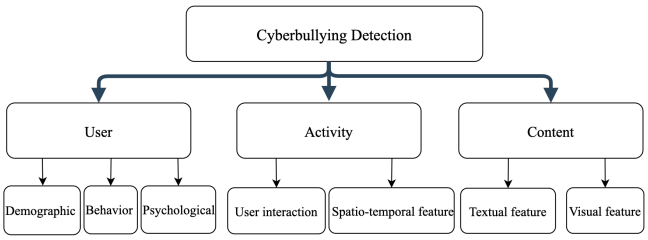


Fig. 5: Feature taxonomy for cyberbullying detection

To find out relationship between user profile and cyberbullying, number of following and followers are integrated as profile feature and the result proves that fake user file is one of the main factors cause cyberbullying because users with fake profiles usually pretend as someone else in order to attack and offend others [25]. User influence in the spreading of cyberbullying is studied to find out how much the contribution the number of followers and friends of a user in Tweeter will be by posting a thread comment by the user. Based on the results, user profile feature is regarded as informative in cyberbullying detection [26]–[28]. Valuable user basic information can be provided by demographic features and using these information can differentiate or compare the behavioral patterns of users based on a group that belong to. Studies have argued that it is common to observe various cyberbullying prevalence rates among different population group [29]. Study points out that sexual orientation is strongly related with cyberbullying victimization and LGBT identification suffers a higher rate of cyberbullying [30].

The most commonly used demographic features are gender, age, race/ethnicity, sexual orientation, socioeconomic status, profession, education level, material status. Dadvar et al [31] investigated the role of a gender-based feature for extracting the cyberbullying on MySpace dataset. Their first observation shows that the frequency of using unpleasant words among female and males are significantly different. For example, female used to express the profanity words indirectly or in the implicit vein. Thus, SVM classifier is used to confirm that the incorporation of gender-based feature can improve the accuracy of cyberbullying detection. A close tie between age and cyberbullying is realized which means people' attitude would change over time. Their meta-analysis method represented that cyberbullying can increase among the boys particularly as they start the high school, and goes down as get older [32]. Gender-based language, age-based language and user location features are added into a semi-supervised learning for cyberbullying detection with a fuzzy SVM algorithm. The evaluation conducted on different scenarios shows the performance improvement of the proposed fuzzy SVM with those three features [33].

B. Behavior features

Behavior features refer to individual user activity mode which contains the information of users' activities in online environment, such as online social network log in frequency, how long user maintains online status as well as user online language pattern. The history of user comments allows

TABLE III: Summary of user-centered features

Feature	Sub-feature	Strength	Weakness
Demographic	Gender, Age, Race(Ethnicity), Socioeconomic status, Education level, Material status	Clear demonstration of user basic info	Noisy, and hard to obtain complete information
Social network profile	Gender, Age, Socioeconomic status, Material status	Indicate user online self-positioning	Cannot guarantee true info
Behavior	Online social network log in frequency&time, Comments & post history, Language pattern	Direct info about user online habit	Vary, and change over time
Psychological feature	Big Five, Dark Triad	Strong indication of user personality	Hard to obtain precise answer

TABLE IV: Popular post content for different user categories [23]

User category	Topic	Hashtag
Normal	uniteblue, feminism, women, tcot, abortion, gender, imwithher, prochoice, womenrights, otrasheffield mtvstars, britney, spears, lana, great, gomez, selena, demi, lovato, antinwonowtina, brexit, voteleave, euref, gamersunite, leaveeu, people, world, voteremain, vote, pushawardsjadines	#mtvstars, #uniteblue, #pushawardslizquens, #pushawardskathniels, #brexit
Spam	porn, tweet, boobs, sexy, pics, vids, tits, antinwonowtina, exposes, erol, love, pushawardskathniels, boobs, retweet, busty, followers, years, girls, again, leaked, lgbt, dino, love, follow, nowplaying, itunes, giveaway, summer, enter, seconds	#boobs, #ass, #porn, #busty, #milf
Bully	feminismisawful, antifeminist, whitegenocide, direction, mtvstars, antifeminism, famous, diversity, hypocrisy, feminista action, offend, crowd, comentario, andreiwi, grollizo, hatebritain, jewfnitedstate, feminista, watchmylifegrow stayandendure, masochist, pigs, feminist, voteremain, paedophiles, genocide, misery, feelthebern, patriarchy	#feminazi, #hateconsumed, #fags, #feminismisawful, #jewfs
Aggressor	zionist, groomed, erol, exposes, jews, promisedlanding, misery, heart, world, necessidade, brexit, leaveeu, more, like, attack, cowards, bluehand, feminismisaw, maga, medical,feminism, venezuela, hatebritain, ormiga, heard, show, abandon, rioux, brad, safe	#gay, #zionist, #feminismisawful, #hate, #brexit

algorithms to monitor user's activities by considering the average of context features like profane words to find out the established language pattern and further to see whether there is any usage of offensive language. Additionally, the posting behavior also allows to trace user's reaction toward the bullying or harassing post in different online platforms [34]. Through the posting behavior, for example, we can observe what would be the real reflection of one whether in YouTube or Facebook when they have experienced a harassing post in the YouTube.

Table IV summarizes popular topics and hashtags posted by four different groups of network users [23], [35]. Normal users prefer to talk about various topics and use different hashtags such as social problems and celebrities while inappropriate content can be found in spammers' post, by which more attentions can be obtained from other online network users and help them to gain followers. Some sensitive issues are more discussed by bully users such as feminism and religion. Unlike normal users expressing their true feelings about popular topics, aggressor users are more tend to deliver negative opinions on those topics.

TABLE V: Big Five features

Feature	Symptoms(Signs)
Extroversion	Talkative, assertive, gregarious
Neuroticism	Emotionally installable, anxious, worrisome, insecure
Agreeableness	Good-natured, forgiving, tolerant
Conscientiousness	Careful, thorough, organized, dependable
Openness	Imaginative, curious, artistic

C. Psychological features

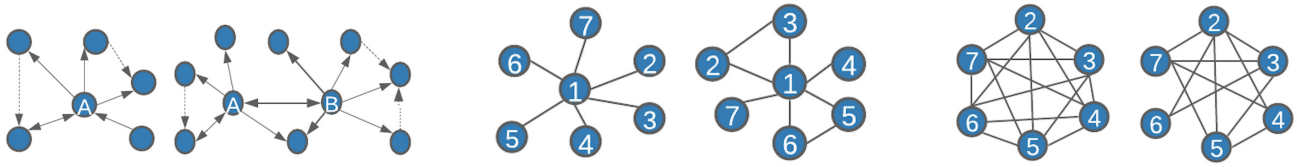
Psychological features are characteristics that define an individual including personality traits and behavioral characteristics. Research has shown that certain groups pf people are more inclined to become perpetrators or victims of cyberbullying violence based on their personality traits. In addition, cyberbullying can threat the victims psychological and physical health. Therefore, some studies have started to extract psychological features to make early intervention or prevention of cyberbullying. One of the most comprehensive approach to recognize personality is based on the Big Five model that describes personality traits from 5 aspects [36], [37]. The Big Five model comes from a statistical study of responses to personality items. Using a technique called factor analysis, researchers can look at people's responses to hundreds of personality items in a test and ask the question "What is the best way to summarize a person?" [38], [39]

As shown in Table V, extroversion defines the tendency of outgoing, sociable, interested in others, decisive, positive, caring more about external events and seeking stimulus. Agreeableness measures the tendency to be kind, friendly, gentle, get along with people, and be enthusiastic about people. Conscientiousness feature indicates how much a person care about others when making a decision. The following is called neuroticism that means the tendency to depression, fear, and moodiness. Openness shows the tendency to be creative, insightful, thoughtful, open-minded, and willing to adjust activities based on new ideas [38], [39].

Another personality model that has gained momentum in cyberbullying research is the Dark Triad, which pays more attention to the shady characteristics of users' personalities. It refers to three different but (to others) unwelcome charac-

TABLE VI: The relationship between Dark Triad and Big Five

Dark Triad/ Big Five	Openness	Conscientiousnes	Extraversion	Agreeableness	Neuroticism
Machiavellianism	-	-	/	-	+
Narcissism	+	-	+	-	/
Psychosis	+	-	+	-	+



(a) 1.5 and Combines ego-network (b) Star and Near-star (c) Clique and Near-clique

Fig. 6: Examples of typical social network graph structures: (a) 1.5 and Combines ego-network; (b) Star and Near-star network; (c) Clique and Near-clique network.

teristics, namely Machiavellianism (that is, lack of empathy and the tendency to engage in impulsive and stimulus-seeking behaviors), psychosis (that is, strategically the tendency to manipulate others) and narcissism (that is, the tendency to feel superior, magnificent, and empowered) [40]. Similar to Big Five model, participants are required to answer a series of question so that researchers are able to calculate their scores [41]. The relationship between Dark Triad and Big Five is presented in Table VI where - means negative correlation between two personalities, + is positive relationship and / is no relationship [40], [42], [43]. Users personality has been linked with cyberbullying with empirical evidence. Cyberbullying detection using the Big Five found that agreeableness and conscientiousness are negatively related to cyberbullies, while extroversion and neuroticism are positively related [44], [45]. Studies exploring the relationship between cyberbullying and darker personalities (Dark Triad) have presented evidence that cyberbullying behavior appears more often with these three traits among which narcissism is found to be more linked with cyberbullying whereas psychopathy is more related to cyber-aggression [46]–[48].

IV. Activity-centered features

Activity centered features intend to characterize user activities in order to understand how cyberbullying users (and cyberbullying content) different from normal users. Such activities are often studied from two main perspectives (1) how users interact with each others, and (2) how same users behave across temporal or spatial scales. For the former, networks are commonly used to study user interactions. For the latter, users' activities are considered spatio-temporal, or time series, for analysis.

A. User Interactions

We define user interaction feature as the activities conducted between online social network users which focuses on group interaction instead of individuals.

1) Local Interaction Feature: For online social networks, node degree which means the number of followers or friends can be a proper scale to explore the person social activity to examine their influence on one another. Users are indicated in a weighted direct graph method to detect the cyberbullying, the researchers consider the volume of interactions or conversation among the users, because they believe that the more rate of connection among nodes goes up, the more the probability of bullying would increase [54]. Similarly, pairwise influence such as edge betweeness and peer pressure are taken into consideration for cyberbullying detection because users can make influence on or be influenced by their peers bullying behavior and even follow those to become cyberbullies. They believe that the influence of a user on others is a function of the weight of social relationships of users on their proposed graph, the time of influence, the probability that influenced user might get involved in cyberbullying, posting content from the influencer, and the social networks of both users [27].

Fig. 6 shows common social network structure graphs. In the 1.5 ego networks (Fig. 6a), 1-ego for direct and 1.5 for dash connection lines; the combined version defined for the relationships of two users (A) as a sender and (B) as a receiver; In the star network (Fig. 6b), the neighbors are fully disconnected, but in the near-star few of them are connected; In the clique (Fig. 6c), all neighbors are fully connected, but near-clique few of them are disconnected.

ADOMS (Anomaly Detection on Multi-layer Social Network) [58] is a multi-layer method using graph mining to detect cyberbullying. The intuition behind this approach is that in the social network, we can consider users showing abnormal behavior regrading their neighbors as outliers. The study attempts to detect the outliers based on scoring the nodes compare to the predefined social networks like Clique/ Near-clique or Star/Near stars to show that to some extended a node can vary from their neighbor nodes. However, since users get involved in various social networks such as Facebook, Twitter, and Instagram, their interactions are considered in different layers through multi-layer networks. In other words,

TABLE VII: Summary of user interaction features

Feature	Cyberbullying implication	Local/Global	Advantage	Disadvantage
Node degree [26], [49]	Total number of immediate contacts	Local	Characterize user popularity	Miss critical points
Average neighbor degree [49], [50]	Mean degree over all of its immediate contacts	Global	Measure homophily	Miss critical points
Edge betweeness [49], [51]	The influence of an edge	Local	Show how important an edge is	All pair shortest path. Computationally expensive
Embeddedness [49], [52]	Mean of the ratio between the set of common contacts and the set of all contacts for the node	Local	Indicate contact frequency between nodes	Could change with time
Tie strength [27], [49]	Number of messages users send	Local	Show how active a user is	High volatility
Cluster coefficient [49], [52]	Measure how well connected the neighborhood of the node is	Global	Show the connection level of a network	Global measurement, no individuality
Number of triangles [49]	measure how intercommunitarian a node is	Local	Indicate strongly knit communities	Relatively expensive in computation
Interaction [27], [53], [54]	Type of interaction	Local	Monitor behavior of individuals	Unable to be generalized
Neighborhood overlap [49], [55]	Overlapped neighborhood	Global	Low complexity. Good performance	Bias towards high degree nodes
Vertex and Edge count [26]	User amount and user connection	Global	Show the components of the network	Does not show the connection between nodes
Diameter [56]	The longest path in a graph	Global	Show readability and separation of network	All pair shortest path. Computationally expensive
Core numbers [57]	The structural embededness of a node in a network	Local	Show node participation in a highly connected neighborhood (k-core)	Computationally expensive to find all k-cores
Contribution index [57]	Contribution frequency	Local	Show how frequently users send/receive message	Could change with time
Jaccard's coefficient [57]	Common neighbors-based	Global	Measure overlapped neighbors	Impacted by cluster coefficient
Weighted node-level metrics [57]	Edge strength	Local	Reflect the strength of social interactions	Could change with time
Tie strength [57]	#messages sent between two nodes	Local	Show interactive level between two nodes	Could change with time
Attention spanning [57]	The amount of direct attention that a node gives to another node,	Local	How much a node is bridging communities	Dependent on the connection level of a node's neighbors
In-degree and incoming messages ratio [57]	Ratio of the incoming messages between a and b to the total incoming messages to b	Local	Show how interactive between a and b	Could change with time
Out-degree to in-degree ratio [57]	Ratio of sent messages to incoming of a node	Local	Show node communication level	Could change with time
Messages to degree ratio [57]	Ratio of two nodes' incoming messages to in–degree ratio	Local	Show a node communication level	Could change with time

an anomaly score in the individual layer is assigned for each node upon its degree similarity to the Clique/ Near-clique or Star/Near stars. The ratio of followers to friends [51] can be computed as an index for users' popularity, total number of their own and liked tweets [32] as another index for activity rate. The studies prove that closeness, betweeness, out-degree centrality respectively can be prominent user interaction features in terms of information that provide for cyberbullying detection in Twitter users.

2) Global Interaction Feature: User global feature represents the community feature in which users are located in. User amount and user connection level can be described as Vertex and Edge Counts. Cluster coefficient measures a node variety relative to the graph density and density corresponds to the ratio of the number of existing edges to the number of edges in a complete graph containing the same number of vertices. The length of the longest path in a graph is described as diameter [56]. Neighborhood Overlap is used to scale the relative position of receiver and sender users in the direct following network as an adjusted version of Jaccard's similarity index [55]. In their research, five related measurements of neighborhood overlap are considered for a given author and target, namely Downward, Upward, Inward, Outward, and Bidirectional. Values of downward and upward measurements indicates the low and high position of sender versus receiver, while inward and outward showing the visibility.

In Table VII, we summarize major features used to characterize user interactions, including their cyberbullying implication, advantage, and disadvantages. The local/global column indicates whether the feature(s) is intended to capture international in a local (a user and his/her surrounding) *vs.* global scale (the whole community).

B. Spatio-Temporal Features

Cyberbullying is normally not an incident occurring only once or in a distinct fashion. Rather, it can be continuous and repeated or persisted over time. Cyberbullying could be

conducted on the same victims for several times and some cyberbullying victims even become the perpetrator of cyberbullying [25], [59]. To date, cyberbullying detection using temporal features still remain lots of potential as there are insufficient studies related to this specific aspect.

In order to model temporal dynamics of cyberbullying sessions, nine temporal features, including *Time to first comment, ICI mean and variance, ICI coefficient of variation, Number of bursts(Commenting behavior trend), and Amount of total activity and its average* , are applied to statistically differentiate the bullying and non-bullying comments [60]. They use a function $\delta(t)$ to denotes the number of comments at a given time t as shown in Eq. (1) where N is the subsequent of comments in time t_i (hours). They also considered the time between any two consecutive comments as inter comment interval (ICI) and $D = \{\Delta | 0 < i \leq N\}$ where $\Delta = t_i - t_{i-1}$ as as a list of N time-deltas. And they computed the level of activity (bursts) at a media session using Poison Surprise method in Eq. (2). A study considers the frequency (repetition) and time of comments posted as temporal features to detect cyberbullying on Instagram dataset [25], [61]. Apart from providing frequency plots in the various interval or over times for comparing the cyberbullying and non-cyberbullying behaviors, they also applied the detection approach developed by [62] to identify the burst activities. According to the results of these methods, when a post shares in the networks (Instagram) whether in the long term or short term, its intensity gradually but not monotonically decreases as time passes.

In a graph temporal model, TGBullying [63], user interaction in the long term is fed into a graph based temporal model to improve the performance of cyberbullying detection. Because cyberbullying is a repetitive action on any social media, the user interaction can pave the for characterizing it based on analyzing both content and temporal as well as tracking the users' roles. In this regard, however, they found the *sparsity* and *characterizing of repetition* and *user characteristics* as strong challenges to modelling the user interaction.

$$C(t) = \sum_{i=0}^{N} \delta(t - t_i) \tag{1}$$

$$A(t) = \sum_{\{t_i | C(t_i) \neq 0, t_i <= t\}} \exp^{-2(t-t_i)} \tag{2}$$

V. Content-centered features

Content centered features are the main category of features used in majority research, where texts and images are the main source of information implying whether a user's message contains bullying content. Features in this category are extremely rich, and majority of them are related to natural language processing and image analysis.

A. Textual features

Textual features are most commonly used to analyze cyberbullying, a lot of research study papers are using textual features as one part of the processes together with classification algorithms such as deep learning to detect cyberbullying from the raw text data [64]. In this paper, we grouped textual feature into linguistic feature, semantic feature and syntactic feature and summarize them in Table VIII.

1) Linguistic Feature: Dictionaries are the predefined set of profanity or hate words in the comments or texts, such as swear words. Nadali [65] reviewed and presented that textual features approach is often categorized as two groups, one is based on a dictionary to filter the cyberbullying posts called "lexical" features. The other group is called "behavioral" features which happen in conversations. However, the performance of these features limited and might not cover all offensive words, in particular those are ended the domain of specific-textual orientations.

Bag-Of-Words (BOW) are a list of negative words such as swear, profane occurring more frequently in any text or document. It is worth to mention that, the word and its order, position in the document do not matter [66]. However, there are some concerns regarding the use of the BOW. It turns out that it is susceptible to the sparsity related the dependency of its features (elements), therefore, it is not able to capture the semantic information [66], [67]. To address this, some works use other features alongside BOW, considering BOW as a single feature might cause miss-classification due to the fact that words can have different usage in the texts. For example, the BoW based method is improved by combining other features such as part of speech, negative connotations with BOW [68].

N-gram is a sequence of n-words or characters allows to count the number of occurrence in size list in the texts. Unlike BoW ignoring the word orders, n-gram can improve the classifier performance thanks to incorporation of some degree in the context for each word [69]. However, n-gram approach is suffering some limitations such that is not efficient for high level distances [70].

Term Frequency Inverse Document Frequency (tf-idf) is the most common numerical statistic feature applied by textual studies that measures the importance of a word in the comments or documents. The tf-idf value increases in proportion to the number of times the word appears in the document, and is offset by the number of documents in the corpus containing the word, which helps to adjust the fact that certain words usually appear more frequently. It performs better than the BoW model with considering the importance of words in the document, but still not able to capture word semantics.

2) Semantic Features: Semantic is an element of a word's denotation or denotative meaning and semantic analysis technology has been considered for cyberbullying detection, such as filtering malicious information and spam in online communication [71], [72]. A novel approach uses word embedding to model words in Tweets [73], so semantics of words is preserved and the feature extraction and selection phases is eliminated. In order to capture semantics, Latent semantic indexing (LSI) is used for cyberbullying detection which is able to bring out the latent semantics in a collection of doc-

uments [74]. LSI is based on Singular Value Decomposition (SVD) which decomposes a term-by-document matrix, A into three matrices: a term-by-dimension matrix, T, a singular-value matrix, S, and a document-by-dimension matrix, D. as shown in Eq. (3).

$$A = TSD^T \tag{3}$$

In addition to above approaches which use latent vector to capture word semantics, some methods use more transparent way to model semantics. Profanity is a lexicon of negative words that is commonly used to detect cyberbullying. Profanity feature together with general features such as tf-idf are categorized and utilized from YouTube dataset to detect the explicit form of abuse verbal by their pattern-based stable patterns [75]. NLP as a supervised classification method is used to detect abusive language over time on the News and Finance comments of Yahoo dataset. In this regard, the researchers use lexicon (Hate speech, Derogatory, and profanity) as a guideline to annotate the data whether the text is clean or abusive from which profanity returns sexual remarks and other negative words [76].

Sentiment analysis based on user social network post reveals users' opinions, emotion and behaviors [23]. Sentiment is usually categorized into the three groups which is positive, negative, or neutral and these can indicate how user feels at the moment they conduct post, comments. Nahar et al propose a graphic based approach to detect predator and victim in the collected dataset [54] where sentiment is selected as desired feature. Regarding the sentiment features and the results prove that sentiment features can improve classifiers performance in cyberbullying detection. For online bullying activities like cyberbullying, or abusive language, text characters are more frequently used by the people, because of their ambiguity property, non-verbal signs [77].

Word embedding is a vector exhibition of words in which words are set with degree of similarity. This vectorized representation of words allow us to capture the discriminative features of dataset like the words distance [76], [78].

3) Syntactic Features: Syntactic features contain typed dependencies and part of speech features extracted from sentences. Part of speech (POS) corresponds words in a text to a specific part of speech such as noun, verb, adjective, and adverbs according to their context and definitions [23]. Chatzakou et al [23] used Tweet NLP's POS tagging library to extract the POS tag from the contexts. Their descriptive statistics represents that the lower rate of adjective as well as adverbs categories are used in cyberbullying comments. POS is used to determine the score of an aggressive text. Typed dependencies means the syntactic grammatical association in the sentence which can be used as features to extract cyberhate in social medias. The Stanford Lexical Parser is used as a popular tool to identify the type dependency. It returns 51 different linguistic labels, namely nsubj, det, dobj,nmod, compound, and advmod. For instance the parser nsubj as an abbreviation of nominal subject provides the relationship of any syntactic subject with other terms or words.

B. Visual features

Visual contents are defined as silent or motion pictures depicting a story or an incident. For example, drawings or depictions can express thoughts, feelings, ideas to be understood by people and by now they have been converted to new forms such as images, videos, or cartoons. Additionally, the rapid and tremendous growth of the technology regarding the telecommunication industry enables users to capture photos, record videos, create animations and to name but a few and instantly shares them with others through online social networks. They have been identified as potential factors in spreading of anti-social movements namely violence, harassment, aggression, and bullying in the online social medias. In order to reduce the pace of such adverse activities, therefore, content of the visual posts need to be described and detected automatically. Feature categories, sub-features and their strengths as well as weakness are summarized in Table IX

1) Image features: Images are considered as one of the key components in visual features, especially on online social platforms which allow the users to share image-based posts. Although images grab user attentions and promote their engagement in these media, they can foster aggression, hate speech, and cyberbullying in these platforms. We summarize the relationship between commonly used image features and cyberbullying in Table X

a) Body posture: Body posture regards the body pose of people in image in which they are pointing a subject such as a gun to someone else. This feature is highly dependent on the presence of a person. Via the cosine similarity that compares the difference between these features regarding the cyberbullying and non-cybebullying images, Vishwamitta et al found that there is a strong correlation between cyberbullying images with body-pose when one takes a front pose and pointing a threatening object towards viewers [82], [98].

b) Facial emotion features: Facial emotion features express one's feelings or emotions in image. In cyberbullying images, predator might bully their victims by showing aggressive or even happy and joyful facial expressions. For example, aggression in image can covey a threatening action to viewers, while happy can indicate mocking [82]. Using K2 algorithm which is a Bayesian method for summarizing a structure from given data helps to learn the structure among facial features. Bayesian Information Criterion determines the score M of K2, which is denoted by Eq. (4), where N denotes the size of the database, D_i is the size of a data, and $P(D_i)$ is the occurrence probability of D_i [107].

$$BIC(\hat{\theta_M}, d) = \sum_{i=1}^{N} \log P(D_i) - d \log N/2 \tag{4}$$

c) Gesture features: Gesture features represent the motions of hand in several poses or gestures that people make in image. Some are not appropriate and even harmful and are used to frighten or mock at the viewers. In the cyberbullying images, the most prevalent gestures can be a loser thumb, a middle finger, a gun gesture, and a thumb down. Based on

TABLE VIII: Summary of textual features and their strength *vs.* weakness for cyberbullying detection

Feature	Sub-features	Strength	Weakness
N-gram [54], [64], [79]–[88]	Unigram, bigram, trigram,char quadgram, skip-gram	Reduce spell variation; Cooperate contexts of words	Less efficient for higher grams and depend on other features
TF-IDF [34], [64], [75], [79], [83], [89]	tf-idf	Word importance is considered	Cannot capture semantics, etc
Bag of words [27], [68], [80], [85]	bag of words	Do not need a predefined dictionary	Has detection issue and dependent on number of list
POS [68], [75], [76], [85], [90]	POS	Capturing word syntactic functions	Has ambiguity
Profanity [76], [89], [91]	Profanity, hate speech, derogatory	Transparent, easy to interpret	Need to define and manage dictionary. Word ambiguity
Word embedding [67], [76], [87], [92], [93]	Word2vec, comment2vec,paragraph2vec	Word as vector for calculation. Flexible dimensionality	Poor interpretability. Need training for new comments
Sentiment [54], [94], [95]	Slang, hashtags	High level semantics. Applicable to words/sentences	May vary in different contexts
Text Character [76], [94], [96], [97]	Emotion, non-verbal sign	Easy to parse and obtain. Simple and accurate	Sparse. Many do not have those signs.

TABLE IX: Summary of visual feature and their strength *vs.* weakness for cyberbullying detection

Categories	Features	Sub-features	Cyberbullying Indication	Strength	Weakness
Image features	Body postures [82], [98]	Front pose, non-front pose	Strong	Show body pose of a people	Dependent on presence of person
	Facial emotion features [82]	Anger, depression, surprise,joy	Medium	Express one's feelings or emotions	Complex to detect
	Gesture features [99]	Middle finger,thumb down, gun gesture	Strong	Express motions of hand or gestures of a people	Need to consider the circumstance
	Object features [82]	Knife, gun	Strong	Show objects in the image	Need to consider the circumstance
	Social features [82]	Anti-LGBT, anti-black	Strong	Represent the usage of ant-social symbols	Complex to detect
	Taxonomy-based features [100]	celebrities, clothes, text, animals, tattoos, sports	Medium	Show the attention attract in the image	Need to consider the circumstance
	Profile-based features [101]	gender, age	Medium	Represent demographic status of people in image	May have false info
	Image type feature [101], [102]	black/white image, color image, line drawing	Weak	Easy to obtain	Poor semantics. Need to consider the circumstance
	Image caption features [100], [103]	Image content labeled on images	Strong	Detailed descriptions, textual content	Limited availability (only apply to news etc.)
	Image captioning features [104]	A verbal representation of images	Strong	Direct interpretation of image content	Noisy, low accuracy
Animation features	Color [105], [106]	Color in the cartoon	Weak	Indication of different types of scene in video clips	Ambiguous for unusual colors
	Local features [105]	Pixels, corners, blobs, sift features	Weak	Finding local structures and objects	Need to explore neighborhood. Computationally expensive

TABLE X: Correlation between image features and cyberbullying

Feature	Attribute	Correlation with cyberbullying
Body pose	Front pose	High
	Non-front pose	Low
Emotion	Joy	High
	Sorrow	low
	Anger	High
	Surprise	Low
Gesture	Hand gesture	High
	No hand gesture	Low
Object	Threatening object	
	No threatening object	High
Social	Anti-LGBT	High
	Anti-black racism	High

the cosine similarity, a study [82] show that there is a strong relationship between hand gesture in images and cyberbullying. Additionally, a study of online involvements [99] on gang violence has shown that the presence of harmful hand gestures is affiliated to the online media images.

d) Object features: Object features regards to the objects in images. As for cyberbullying, however, some people post images containing the objects to threaten or intimidate the viewers. The most common attributes of which can be gun, knife, revolver, and so on. Most studies explored the presence such objects in cyberbullying. For instance, a significant correlation between the cyberbullying images and such objects in mage is found [82].

e) Social features: Social features represent the usage of ant-social symbols in images. In cyberbullying images, perpetrator uses these symbols to demean and offend special groups. Since the range of this factor is very vast, it is so difficult to define any specific dictionary or attributes.

So, each study should limit itself to the images containing these symbols in their dataset. The presence of race or adult content in images is considered as one of visual features to detect cyberbullying in a study [101]. Vishwamitta et al [82] define the presence of anti-black racism as well as anti-LGBT symbols as a sign of cyberbullying.

f) Taxonomy features: Taxonomy features define the special object in image that can grab the attention of the viewers. They can also happen in the absence of person. Money, drugs, animals, and even celebrities are sample attributes of this feature can convey an intent of cyberbullying [100].

g) Profile based features: Profile based features represent the demographic status of people in image such as gender or age. Some studies [101] found that the gender of people in image can lead to the cyberbullying.

h) Image type features: Image type features define the color of images. Images with black or white colors as well as clip art and line drawing type are considered for cyberbullying detection [101]. Colors like red, green, blue (RGB) can be helpful in identifying bullying and non-bullying images [102].

i) Image caption features: Image captions are any textual content labeled on images in order to express the image. Studies have explored the role of the caption in cyberbullying involvements and proposed the approaches to extract the main topic of the captions. Caption feature is incorporated in the study for hate detection. The researchers employ Google Vision API Text Detection model to extract the texts from images. Then, these texts are inputted the model to score the relationship between the caption and the area of image where they are appeared [103] and Hossienmardi [100] also the relationship between the presence of captions and cyberbullying is also explored based on Instagram dataset.

j) Image captioning features: Image-generated features are obtained from the converting the contents or objects of images into the words or sentences. In other words, it provides a verbal representation of images. However, the downside of these features is that the lots of information might be lost in this process. To address the issue, therefore, the attention mechanism proposed where the salient region of an image calculated. To calculate the attention the encoder-decoder architecture is used to translate images to sentences [104]. In the encoder, the image is encoded to a sequence of words. It generates a caption vector like $v = \{v_1, v_2, \ldots, v_C\}$, where v_i is in R^K, and *K* and *C* are the size of vocabulary and the length of the caption respectively. Then the annotation vector like $a = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_L\}$ where each $\mathbf{a}_i$ is in R^D and related to each part of the image produced by CNN. Image captioning features are denoted as $X_1 = \{I_1, I_2, \ldots, I_i\}$ and text features are presented as $X_2 = \{T_1, T_2, \ldots, T_i\}$ in VisualBert architecture in which a image like *I* is extracted in the multiple region features $f = \{f_1, f_2, \ldots, f_n\}$ by usinf Faster R-CNN. Then it is converted into a visual embedded e_v by Eq. (5), where e_s indicates the input is image or text. Similarly, text inputs are embedded as shown in Eq. (6), where f_t and e_p are token embedding and positional embedding (relative position) for each token in the sentence. Finally, it outputs two multi-model representation $(t_1, t_2, t_3, ...)$ [108]. Likewise, a vector of captions like $X_3 = \{C_1, C_2, \ldots, C_i\}$ are extracted by the Image Caption model based on the attention mechanism. This vector is fed to the BERT architecture (Language model). It outputs a textual representation $(p_1, p_2, p_3, ...)$ having same dimension with VisualBert outputs. Through the concatenation or bilinear transformation, their information is transmitted into one vector [109].

$$e_v = f + e_s \tag{5}$$

$$e_t = f_t + e_s + e_p \tag{6}$$

2) Anime & Cartoon: Cartoons are videos depicting stories or occurrences in imaginary and semi-realistic ways. They have been an integral part of a wide spectrum of people especially young generation and they spend fairly a large amount of time watching these anime movies. Apart from entertaining facet, they are considered as an informative resource that fosters its audience's awareness. By considering of all benefits of animations, however, today they are subject of huge controversies and debates. Recent studies show that the impact of the aggressive and offensive contents spreading by cartoons is not limited within the real life and can encourage the detrimental activities, namely hate speech and cyberbullying, and cyber-aggression in online social platforms.

a) Color: As a low-level feature in this literature, colors can be an indication of different types of scene in video clips. For example, the distribution of violent scenes is more likely different that of non-violent in any movie. In general, the global color histogram provides all information of colors of visual feature as a histogram to show how they distributed over different bins. Color histogram is applied to extract the violence and non-violence activities in the cartoons [105], [106]. For image color feature extraction, Fisher kernels are introduced to characterize the dataset samples. Its gradient vector is shown in Eq. (7) and the information matrix is illustrated in Eq. (8). Fisher kernel is a powerful framework and it combines the advantages of pattern classification generation method and discrimination method. The idea is to use a gradient vector to characterize a signal Probability Density Function Modeling the Signal Generation Process (pdf) [110].

$$\bigtriangledown_\lambda logp(X|\lambda) \tag{7}$$

$$F_\lambda = E_X[\bigtriangledown_\lambda logp(X|\lambda) \bigtriangledown_\lambda logp(X|\lambda)'] \tag{8}$$

b) Local feature: Local features focus on the specific or exciting regions of a given visual frame. In local features, a patch of a given frame should differ from their immediate surroundings based on the color, texture, shape, etc. Corners, pixels, and blobs can be simple examples of the local features. Thanks to this property, they are usually utilized for the object detection in videos and images and are applied for cyberbullying detection [105].

VI. Feature Interaction and Platform

A. UAC Feature Interaction

User-Activity-Content triangle illustrates the interconnection between three type of features. Following the review of features with respect to each vertex of the UAC triangle, we summarize the inter and intra feature interactions between users, activities, and content in Fig. 7. Because UAC futures focus on different aspects, we use two major facets, semantic indication *vs.* cyberbullying indication, to project them to Fig. 7. More specifically, semantic indication denotes the degree of semantics the feature may imply, and cyberbullying indication represents the strength of correlation that they feature may reflect during a cyberbullying incident.

For all summarized features, texts are most direct and accurate in capturing semantics because languages are the main form of communications. Visual perception is the second most important source of information acquisition, but understanding semantics of visual objects relies on computer vision and image processing, which are often less accurate in comprehending semantics than texts. Collectively, textual and visual features are the most effective ones in capturing cyberbullying indication, mainly because of their strength of semantic relevance. In addition, content-centered features are much easier to harvest, because majority cyberbullying or cybervoilence actions are carried out through texts, images, and videos etc.

Activity centered features can still reflect semantics and shed light for cyberbullying indication. For example, constantly sending messages to a receiver is considered a harassment behavior, especially if the message has no meaning or has a negative sentiment. A person repetitively posting and commenting on pages with aggressive content implies a strong cyberbullying indication. In this case, users' activities need to be analyzed by social network analysis algorithms [111], in combination with content information.

User centered features are considered less informative in terms of semantic indication and cyberbullying indication. This is mainly because of two reasons: (1) user features are always noisy and inaccurate, containing many missing or incorrect entries; and (2) user personality and characteristics are compounded factors, making it difficult to capture their semantics. Attacker may disguise themselves using very positive profiles, but their actions, and content associated to the actions, will eventually expose their true cyberbullying indication.

Overall, content features have the strongest semantic indication and cyberbullying indication, in which textual features express a more clear indication than visual features. In contrast, demographic, behavior and psychological features from users are inferior to Content in both aspects, in which behavior shows the best in cyberbullying indication. Behavior trend features from Activity has the strongest cyberbullying indication which surpasses local interaction, global interaction as well as spatio-temporal features.

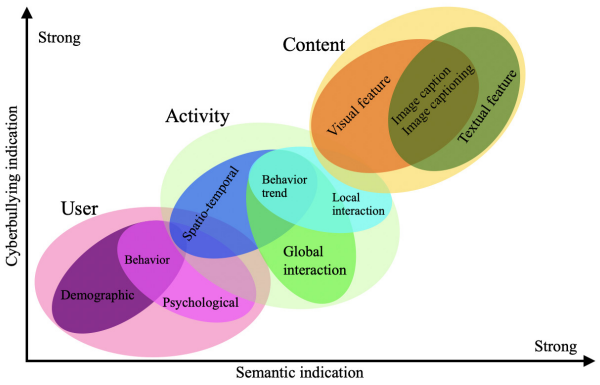


Fig. 7: Inter-feature and intra-feature interactions of the UAC triangle vertices with respect to semantic indication ($x-$axis) and cyberbullying indication ($y-$axis). Features towards the origin indicate weaker semantic and weaker cyberbully indication.

B. UAC Features and Platforms

Features selected from different platforms may vary in terms of availability and quality. For example, users from LinkedIn are more likely to use their real names and provide more personal information. Instagram posts are more image-based. In Fig. 8 we discuss the common selection of features in UAC triangle in terms of 23 common online platforms subjecting to cyberbullying or cyberviolence. Although majority of selected platforms have social network components, some systems, such as MitBBs, craiglist, SMS, do not have direct social networking functionality.

For multimodality-based platforms, such as Instagram and TikTok, studies are more focused on their content features. Multimodality is the interaction between different representational modalities, such as the interaction between textual expressions and videos/images. Features extracted from the majority platforms are content features and user activity features, which makes sense because that is what most people do on online social networks, posting messages/images/videos or interacting with other users through comments, etc. However, for platforms like Linkedin, user-centered and activity-centered features are more important and easier to get such as demographic feature.

VII. Cyberbullying Detection Methods

Following the user-activity-content triangular view, we now review machine learning methods for cyberbullying detection. Majority cyberbullying detection methods are based on supervised learning, in which bullying and non-bullying episodes in online social platforms are differentiated. They are usually classified into supervised learning methods and weakly supervised learning methods.

A. Supervised Learning Methods

Numerous supervised machine learning algorithms have been applied for bullying identification in the virtual social media such SVM, Navie Bayes, Random Forest, Logistic,

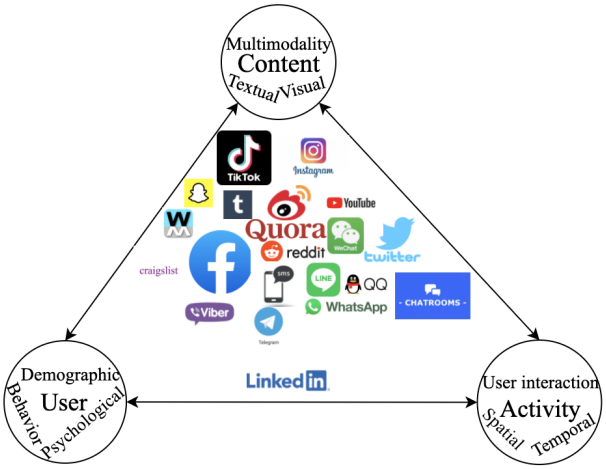


Fig. 8: A summary of common online platforms subjecting to cyberbullying or cyberviolence, and their features in terms of the UAC triangle

JRip, J48, and to name but a few. It is also worth to mention that the performance of supervised learning dependent upon various factors, in particular the dataset, data labeling, and features selection. Furthermore, most supervised methods in cyberbullying studies have been concentrate on textual features, while some have recently confirmed that considering other features like social networks or temporal can improve the models performance. We will demonstrate supervised learning methods from three aspects.

1) Text-based Method: Simple feature based methods are the most common approach for cyberbullying detection. Vast majority of studies have been attempting to identify bullying incidents from textual contents such as comments and posts from online social medias. Supervised machine learning combining with a bunch of the textual features, *e.g.*, BoW, n-grams, TF-IDF and word2vector have been exploited to do the classification among bullying and non-bullying incidents, among which SVM is the most frequently used classification method, and Naive classifiers are the most accurate approaches [112]–[114].

While simple, one major weakness of text based methods is the low accuracy. Most of the text based methods analyze textual features which contain aggressive words, however, under some circumstances, aggressive words do not represent cyberbullying and some real cyberbullying incidents may not contain aggressive words. In addition to the sparsity issue, in many cases, cyberbullying is implied in the sentiment of sentences, pictures, or are tied to the behavior of the senders. Therefore, other types of methods, such as network-based, temporal pattern based approaches are nice complement to simple text-based methods.

2) Network-based Method: Social network relationship features are taken into consideration for cyberbullying detection. The use of social network features such as user activities or behaviors, the aggressive level of a social network group in any online social media are combined with other textual features. Node2vec is used to analyze social network relationship features. The nodes in the graph indicates every user in social network and their labels indicate whether the user belongs to a cyberbullying group. Different single classifier such as Random Forest, J48, Naive Bayes, SMO, Bagging, ZeroR or the combination of them have been applied to apply the extracted information [26], [115]. C4.5 decision tree is considered as a popular tool because it can support both discrete (categorical) and continuous features such as social network features, user demographics [27]. Recently, the integrated supervised learning methods such as Ensemble classifiers as well as Fusion approach are becoming popular for cyberbullying detection, because this kind of methods can support using multiple features like network-based features along with other types of features [23], [116]. Ensemble classifiers is upon on a set of multiple classifiers in which their single decisions (weighted or non-weighted) are combined to classify new data which is able to produce a better performance compared to individual classifiers since the errors of each classifier is eliminated by averaging over the decisions of multiple classifiers.

3) Temporal-based Method: The last method is temporal based detection methods. To detect cyberbullying incidents as early as possible, features measuring time difference of the consecutive comments are applied with supervised learning methods. Two different specific early detection models, named threshold and dual are proposed. In threshold method, the decision function for the threshold model is formatted as Eq. (9), where m can be any machine learning model like Random Forest or Extra tree training the base line features, $th_+()$ and $th_-()$ are thresholds for negative and positive cases. In this function, the final decision is made based on if there are enough evidence determined by positive and negative class probabilities or not.

$$\delta_1 = (m, th_+(), th_-()) \tag{9}$$

Regarding the dual method, in contrast, two different machine learning models are trained to detect positive and negative cases separately as well as independently as Eq. (10), where m+ and m- are positive and negative learning models respectively [117], [118]. These two independent models are trained with an independent set of features separately where one is to detect positive classes and the other detects negative cases.

$$\delta_2 = (m_+, m_-, th_+(), th_-()) \tag{10}$$

B. Weakly Supervised Learning Methods

Considering that labeling a large dataset is not only time demanding, but also needs a massive budget allocated to cover the expenses, recently, weakly supervised learning models (WSL) have been proposed in which a completely annotated training dataset is no longer needed. In weakly supervised learning, there are three main approaches that can lessen the heavy load of the data labeling process, namely incomplete, inexact, and inaccurate.

1) Incomplete Process: The incomplete process only allows a small subset of instances being labeled. Inexact labels allow multiple instances to share labels (*e.g.*, multi-instances labels). Inaccurate data are labeled like the strong supervision, but there might be label errors in the data [119].

Suppose that we have a binary classification problem considering two *Y* and *N* classes, and the task is the learning of $f : \mathcal{X} \rightarrow \mathcal{Y}$ from a training data set like $\mathcal{D}$. With a strong supervision, an annotated dataset should be a set like $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $\mathcal{X}$ is a feature space, $\mathcal{Y} = \{Y, N\}$, $x_i \in \mathcal{X}$, and $y_i \in \mathcal{Y}$. In order to manage the unlabeled dataset, several of approaches have been proposed such as active and semi-supervised learning. Active-based learning allows human intervention to annotate selected unlabeled instances with respect to the anticipated costs, while semi-supervised method tries to take advantage of them without human expert, in which the selected unlabeled data can be treated as a test data(transductive) or the test data are considered unknown while unlabeled instances are not selected as test data (pure).

For pure semi-supervised learning, the question is how the unlabeled instances can get involved in the prediction process to enhance the related model. Recently, some studies have attempted to answer this question by introducing innovative algorithms such as augmented training [120]. In augmented training, the primary training dataset like T_n is initially divided into two groups such as Y_n as well as N_n, that is, $T_n = Y_n \cup N_n$, and U_n denotes the unlabeled data.

2) Inexact Process: In inexact supervision part, the dataset is partitioned into annotated subsets of features instead of the individual features. This approach is also called multi-instances learning (MIL). Suppose $D = (X_1, y_1), \ldots, (X_m, y_m)$ where each X_i is called a bag so that $X_i = \{x_{i1}, \ldots, x_{im_i}\} \subseteq \mathcal{X}$, and $x_{ij} \in \mathcal{X}$ $(j \in \{1, 2, ..., m_i\})$ and m_i is the number of instances in X_i. X_i is a positive bag *i.e.* $y_i = Y$ when there is a x_{ip} so that $p \in 1, 2, .., m_i$. In other words, the standard assumption in MIL is that observing just one positive instance in a given bag can lead to be annotated as positive.

3) Inaccurate Process: For inaccurate supervision, the training data set is similar to the strong supervision, but the value of y_i may be incorrect, because of annotator errors. Recently, it has been utilized for cyberbullying detection on online social medias where labeling the whole data is usually considered as a tedious and time-consuming activity via the traditional learning methods. Among those approaches, participant-vocabulary consistency (PVC) [121] and co-trained ensemble models (CEM) [122] are more popular. PVC attempts to extract cyberbullying involvements in the online platform through evaluating users roles (bullier or victim) and the bullied vocabulary are used in their conversation in the same time. This is because that the bully language strongly depends upon the user role or interactions in a given bully activity.

Denote *U* and *M* a set of users and massages in a given online platform respectively, and *s(m)* and *r(m)* represent a message like *m* sent from user *s* to *r*. To measure the contribution of users and vocab, u_i, b_i, v_i denote the contribution of a user *i* in any bully activity, their bully and victim scores respectively. The integrated scores of participant and the average of vocab help to predict the bully score of any interaction as shown in Eq. (11), where $f(m) = \{x_k, \ldots, xl\}$ is a set of n-gram features describing the message *m*, and k represents the bully score of a given vocab due to the presence of its corresponding feature f_k. Contrast to PVC, in CEM, the detectors like key phrases and two user-based bully and victim tendency scores are replaced with ensemble of two rich learners that co-train each other. One learner checks the messages for searching bully language, while the other relies on the social structure. In the first learner, a single message is fed to the classifier and it outputs a score of bully or harassment, *i.e.*, $f : M \rightarrow R$. However, for the second learner, outputs are an ordered pair of users, that is, sender and receiver. So, the output is a score showing the sender bullies the receiver or not, *i.e.*, $g : U^2 \rightarrow R$. With respect to these scores, the main goal of training is minimizing model in Eq. (12) based on the parameter space (Θ), where the first function is consistency loss that evaluates the disagreement between message and user classifiers.The second one is a weakly supervision loss that lies annotated key phrases of harassment messages like indicator and counter-indicator.

$$(b_{s(m)} + v_{r(m)}) + \frac{1}{|f(m)|} \sum_{k \in f(m)} w_k \tag{11}$$

$$min_{\Theta} \frac{1}{2|M|} \sum_{m \in M} (f(m; \Theta) - g(s(m), t(m); \Theta))^2 + \frac{1}{|M|} \sum_{m \in M} l(f(m); \Theta) \tag{12}$$

C. Unsupervised Learning Methods

The main shortcoming of supervised learning methods is the labeling of huge and high-dimensional data, because it is time demanding as well as expensive. Even though weakly supervised approaches attempt to alleviate this issue, they still depend upon the labeling of data. Some studies have switched to exploit unsupervised approaches that mainly based on pattern mining. In general, unsupervised learning methods lies on searching the patterns in the data that share similarities. To our knowledge, several associated methods have been used regarding cyberbullying detection, mapping based method and outlier detection method.

1) Mapping Based Method: As an unsupervised method, Self-organizing Mapping (SOM) attempts to project a big dataset into a lower-dimension space of neurons in order to discover any similar patterns [123]. In other words, the dataset mapped on one-single layer of a linear 2D with a fixed number of neurons (units) such that the topology of neurons on the layer can be as a rectangular or hexagonal network. Data is categorized with the same attributes by searching their corresponded neurons' layer. With the required input $x = \{x_1, x_2, .., x_n\}$ and model vectors $w_i =< w_{ij} >$, $w_i \in R^n$, the best matching (winner) should be determined

as shown in Eq. (14), where t is a current training iteration. Through training process, learning rate $\alpha(t)$ decreasing with time in Eq. (15) to reduce the difference between model vector and their corresponding input patterns and

$$\beta_{c_i}(t) = exp(-\frac{||r_c - r_i||^2}{2\delta(t)^2}) \tag{13}$$

is a neighbourhood function. When it comes to cyberbullying detection, although SOM is considered as a powerful tool in structure detection for any large dataset, it is neither inherently hierarchical nor random process to provide a distribution of data. Therefore, GHSOM is applied to work with social medias to improve the accuracy [124]. GHSOM has a dynamically growing structure which allows the distribution of data presented on hierarchical multi-layers where each layer is working as an independent SOM [125]. The difference between the input vector and model vector of units are checked by Mean Quantization Error (MQE) as indicated in Eq. (16). It measures the heterogeneity of projected input data on the unit where x_j and w_i are the input and model vectors respectively, C_i is a universe set of all input vectors. The fine-tuned GHSOM model is tested again Twitter, YouTube and Formspring and is proved to be more accurate in cyberbullying detection [124].

$$c(t) = argmin_i\{||x(t) - w_i(t)||\} \tag{14}$$

$$w_i(t+1) = w_i(t) + \alpha(t) \cdot \beta_{c_i(t)} \cdot [x(t) - w_i(t)] \tag{15}$$

$$mqe_0 = \frac{1}{n_l}\sum_{x_j \in I} ||w_0 - x_j||; \quad n_C = |C_i| \ and \ C_i \neq \Phi \tag{16}$$

2) Outlier Detection Method: Most studies in unsupervised learning are based on pattern discovery from considerable portion of dataset, while there are some applications in which detecting exception or anomalous cases can be more interesting or useful than the common cases. Local Outlier Factor (LOF) is a density-based outlier detection technique and considers multivariate outlier detection that can be efficiently used for low-dimensional datasets [126] and it has been employed in cyberbullying detection. LOF assigns anomaly scores to data points and the anomaly score for node i in layer l is computed in Eq. (17) [58]. Based on this technique, a methodology called ADOMS (Anomaly Detection On Multilayer Social networks) is proposed and experimental results on several real-world multi-layer network data sets show that this method can effectively detect cyberbullying in multi-layer social networks.

$$aScore_i^l = LOF(E_i^l, N_i^l) \tag{17}$$

D. Time Series Method

The aforementioned methods and algorithms have been mainly focused on the detection task, and view of them try to predict the level or severity of bullying for the future state in terms of the current knowledge. For cyberbullying, apart from detecting it, another important issue is to control and monitor the trend of cyberbullying incident in each platform overtime. To address the above issues, some studies have attempted to take advantage of times series capacities in forecasting [84] and controlling the events overtime [127]. As for cyberbullying detection, the main tendency is based on using dimension reduction in times series data to capture the linguistic behavior of users [127].

1) Dynamic Time Warping Method: Dynamic time warping is an algorithm which is designed to align two sequences by warping the time axis to find an optimal match [128]. In the research, the similarity between two time series is found by Dynamic Time Warping algorithm which calculates an optimal warping path between two series of time series [129]. For forecasting time series, Multi-Layer Perceptron (MLP) structure is employed with the data processing function in Eq. (18), where m is the dimension of inputs, n denotes the number of inputs, w_i represent the MLP hidden layer coefficients, u_j are the output layer coefficients, h denotes the number of nodes in the hidden layer, and $\Phi(x)$ are output and hidden activation functions [84]. The combination of Dynamic Time Warping algorithm and the proposed MLP can provide a instant indicators of the severity of cyberbullying.

$$Y_t^* = \Phi_{output}(\sum_{j=1}^{h} \Phi_{hidden}(\sum_{l=1}^{K}\sum_{i=1}^{m} w_{m(l-1)+i} X_{it-1} + w_0) + u_0) \tag{18}$$

2) Temporal Pattern Model: Temporal features are incorporated in modeling cyberbullying behavior at comment-level to facilitate cyberbullying detection [130]. In the paper, two methods are introduced for modeling the temporal patterns of cyberbullying behaviors. The first one is called HANCD for Time Interval Prediction [131]. According previous research, the majority of bullying comments appears in the first several hours after users' original posts and a short time interval can be found between consecutive cyberbullying comments [61], [132]. HANCD is applied to predict time intervals to utilize these aforementioned patterns. Its objective function is described in Eq. (19), where β_1 and β_2 are the parameter that balance cyberbullying detection task and time interval prediction task separately, and ℓ_1 and ℓ_2 are the first and second objective function.

$$\ell = \beta_1\ell_1 + \beta_2\ell_2 \tag{19}$$

Different with HANCD, the second method HANT implements temporal encoding to model the comments structural information based on their posted time. The encoding function is shown in Eq. (20). In the equation, t_j is the timestamp of comment j in a social media session with corresponding encoding p_{t_j} and the w_k is the angular frequency.

$$p_{t_j}^{(i)} = g(t_j)^{(i)} := \begin{cases} sin(w_k \cdot t_j), & \text{if } i = 2k \\ cos(w_k \cdot t_j), & \text{if } i = 2k+1 \end{cases} \tag{20}$$

3) Trend Analysis: To understand how events such as COVID-19 can impact cyberbullying involvements, a statistical approach using change point analysis is applied. Assume let $X_1, X_2, ..., X_T$ are T sequential observations, and suppose there is at most one change point location τ in the mean as indicated in Eq. (21) where δ is a constant. If the mean of random variable is different before and after in the given location, it will be considered as a change point. According to this

research, the majority of changes in time and location can be attributed to COVID-19 and this confirms that cyberbullying is increasing among Twitter users [127].

$$E(X_i) = \mu, \quad if\ i \leq \tau \quad and\ E(X_i) = \mu + \delta \quad if\ i \geq \tau \tag{21}$$

E. Transfer Learning Method

For cyberbullying detection, the lack of labeled dataset has become one main challenge, therefore, transfer learning is considered as a sophisticated and economical technique to address this issue [133]. Unlike supervised and semi-supervised approaches assuming the distribution of both labeled and unlabeled data should be the same, the label distribution of both tasks ca be completely diffident but related in transfer learning. Transfer learning is based on two main components, domain and task. Domain includes two parts, χ is the feature space and $P(X)$ denotes the marginal probability in which $X = (x_1, x_2, ..., x_n)$ is a vector of instances. Similarly, the task is upon two components, $\mathcal{Y}$ denotes the label space and $f(.)$ is an objective function from training data that predicts the related label. Therefore, both domain and task can be notated as $D = \{\chi, p(X)\}$ and $\tau = \{\mathcal{Y}, f(.)\}$ respectively.

Suppose there are two given concepts such as S and T as source and target with D_S and D_T respectively. The transfer learning attempts to enhance the prediction function of T, $f_T(.)$ by applying knowledge of D_S and τ_s, where $D_S \neq D_T$, or $\tau_S \neq \tau_T$. It has been applied in cyberbullying and hate speech detection on online social medias. Two different datasets (t_1 as source and t_2 as target) are fed into the deep neural network where both pass through the shared task like pre-processing (word token), ELMo embedding (word representation) bi-directional LSTM (BLSTM), max-pooling, but split in the classification phase. The main motivation of using transfer learning in this investigation is that the learning of t_1 can help to improve the classification of hate speech of the target tweets t_2 [134]. Their transfer learning model is capable to leverage several smaller, unrelated data sets to embed the meaning of generic hate speech and achieves the prediction accuracy with macro-averaged F1 from 72% to 78% in detection tasks.

Three different methods of transfer learning are checked to see how the knowledge gained from applying BLSTM with attention models in one dataset can improve the cyberbullying detection in other different datasets. The first is Complete Transfer Learning (TL1), where a trained model from one data set is directly used to detect cyberbullying in other data sets without any additional training. Next is Feature Level Transfer Learning (TL2) and only the information corresponding to the features (word embedding) learned by a model are trained by other datasets. The last one is Model Level Transfer Learning (TL3), in which the trained model on one data set and only the learned word embeddings are transferred to another data set to train a new model. These three transfer learning models outperform other machine learning methods on Formspring, Twitter and Wikipedia datasets [135].

F. Deep Learning Method

Currently, classical deep learning methods such as CNN, LSTM, BLSTM, and BLSTM have started to contribute in cyberbullying detection. The main advantage of these methods compared with other conventional machine learning approaches is that they do not need feature engineering, but only need the data embedded in the vectors as input to be fed into the algorithms. Since data in cyberbullying detection are mainly textual, social network and visual based contents, and each type might have taken different approaches to be represented as a vector, thus, in this section we explore the current methods proposed for representing those features in Deep Neural Networks (DNN). Before being fed into DNN, data should be reconstructed. For example, textual contents need to be encoded to word vectors, while network information should be reconstructed as a graph.

1) Textual Feature Representation: There exists an embedding layer in DNN models to process a fixed sized sequence of words in which each word is encoded as a real-value of a vector. Couple approaches have been identified to set embedding the words, GloVe, SSWE [135]. Global Vectors for Word representation (GloVe) is a model to embed words based on their global statistics in a corpus in which the statistics of words co-occurring containing important information can generate meaning [136]. In GloVe, X_{ij} is an entry of the matrix X denoting the number of times word X_j occurs in the context including X_i and $\sum_k X_{ik}$ represents the number of times every word appears in the context of X_i. The weighted least square is shown in Eq. (22), where V denotes the size of vocabulary. Similar to GloVe, there is another method considering the only syntactic context of words, C&W [137] where an original ngram and its corrupted version are fed into a neural network to assess the embedded words. The corrupted ngram is provided by replacing a random word of the original word. The neural network includes four layers, $Lookup \rightarrow Linear \rightarrow hTanh \rightarrow Linear$. It outputs the score of original and corrupted ngrams as $f^{cw}(t) = w_2(a)+b_2$ where w_1, w_2, a, b are the parameters of linear layers. The loss function defined in Eq. (23) is to make sure the score of the original n-gram is the least.

$$\hat{J} = \sum_{i,j=1}^{V} f(x_{ij})(w_i^T \tilde{w_k} - log(X_{ik}))^2 \tag{22}$$

$$loss_{cw}(t, t^r) = max(0, 1 - f^{cw}(t) + f^{cw}(t^r) \tag{23}$$

Glove and C&W may not successfully capture the sentiment information of contexts. In this regard, Sentiment-Specific Word Embedding (SSWE) attempts to address this issue through integrating both syntactic and sentiment information [138]. To do so, two phases models need to be set up to provide the score of the text as well as its polarity. After generating model score through comparing the original and corrupted ngrams by C&W. To learn the polarity of text through the neural network-based approaches, more layers need to be added on C&W model. To understand the polarity

of embedded words, a softmax (conditional probability) layer added to the last linear of C&W model to annotate the text. This phase of method is denoted as $SSWE_h$ and its loss calculated by cross-entropy as shown in Eq. (24), where K denotes the number of sentiment polarity labels and $f^g(t)$ ($\sum_k f^g(t) = 1$) and $f^h(t)$ represent the gold sentiment distribution and predicted sentiment respectively. $SSWE_h$ does not depend upon corrupted ngram, but its limitations are too strict and need to be relaxed. The relaxed model is referred to the $SSWE_r$, and its loss function is replaced with $\mathbf{f}_0^r$ and $\mathbf{f}_1^r$ as positive and negative predicted scores in Eq. (25), where $\delta_s(t)$ is an indicator function showing the sentiment polarity. In some cases, C&W model and $SSWE_r$ are concatenated to develop a new model, $SSWE_u$ to capture language model score and sentiment score from sentences respectively. In this case, the loss function of $SSWE_u$ should be a combination of $SSWE_r$ and $C\&W$ as in Eq. (26), where $loss_{cw}$ denotes the loss function of $C\&W$ and $loss_{us}$ is the loss function for the sentiment polarity.

$$loss_h(t) = -\sum_{K=\{0,1\}} f_k^g(t) \cdot log(f_k^h(t)) \tag{24}$$

$$loss_r(t) = max(0, 1 - \delta_s(t) f_0^r(t) + \delta_s(t) f_1^r(t)) \tag{25}$$

$$loss_u(t, t^r) = \alpha \cdot loss_{cw}(t, t^r) + (1 - \alpha) \cdot loss_{us}(t, t^r) \tag{26}$$

2) Network Feature Representation: Complex social network relationships are usually represented in graph structure. Graph embedding is commonly used to convert the graph into low dimensional vectors. Graph Auto Encoder (GAE) [139] has been known as a powerful tool, which follows the encoded-decoded approach where a given social network is first encoded in the low-dimension representation and then reconstructed in the decoded step. Assume $G = (V, E)$ with $U = |V|$ users is a given social network, $A \in R^{U \times U}$ and $X \in R^{U \times D}$ are adjacency and feature matrices to represent the input graph as well as features of nodes in the input graph respectively. These matrices are first taken by Graph Convolutional Network (GCN) to generate a latent matrix Z defined in Eq. (27), where σ is the logistic sigmoid function. Decoder reconstructs adjacency matrix A by an inner product between latent variables to minimize error in Eq. (28). The second popular way to convert the graph is Node2vec, in which the original graph structures and characteristics is able to be preserved and the transformed vectors will be similar if nodes have similarities. Node2vec uses a random walk to explore the neighbour nodes as shown in Fig 9. Because of the combination of Breadth-first search and Depth-first search, both time and space complexity is optimized and the efficiency is also improved [114], [140].

$$Z = \mu + \sigma * \epsilon; \quad \epsilon \sim N(0, 1) \tag{27}$$

$$\begin{aligned} g &= \frac{1}{2}||A - \hat{A}|| \\ \hat{A} &= \sigma(ZZ^T) \end{aligned} \tag{28}$$

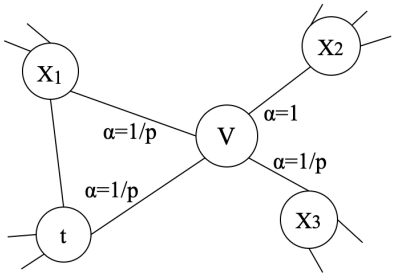


Fig. 9: Random walk for Node2vec: When walking from node t to v, bias α is introduced for each step. The next step is decided by evaluation of the transition probabilities on edges(v,x) determined by return parameter p and in-out parameter q.

3) Image Feature Representation: When it comes to dealing with unstructured data, especially image data, deep learning models are preferred. For online social networks, images can be found everywhere from users' posts to their comments. Fine-Tuned VGG-16 is one deep learning method employed for extracting image features to identify cyberbullying [141], [142], which takes an image with spacial dimension and generates a vector To do so, it is constructed from numbers of conventional layers and dense layers. For each hidden layer, ReLU is used as activation function. Recently, genetic algorithm (GA) has been used to optimize the extracted images from VGG-16 to identify cyberbullying detection. It begins with the random initial selection of the population and then combined by the operators like cross-over to expand the size of current population via generating off-springs and mutation that is used to increase the variation in each generation [142]. The process to convert images into vector for further cyberbullying detection is shown in Fig.10 ([82], [143]), in which after converting to vectors, images with cyberbullying meaning are closer to each other while non-cyberbullying image is far from cyberbullying ones.

G. Method Summary and Comparison

In Fig. 11, we summarize and compare different methods, with respect to the UAC triangle, which demonstrates the focus of different methods in using features for learning.

To compare the requirements of methods used for cyberbullying detection, we summarize their dependency to labels, features, and their adaptability in Table XI, where a symbol $+$ means positive correlation whereas a symbol $-$ denotes a negative correlation (the number of +/- denotes the degree of correlation). For example, labels are required for supervised learning methods (+++), but not required for unsupervised learning methods (-).

To further compare the strength and weakness of cyberbullying detection methods, we summarize their advantages and disadvantages in Table XII.

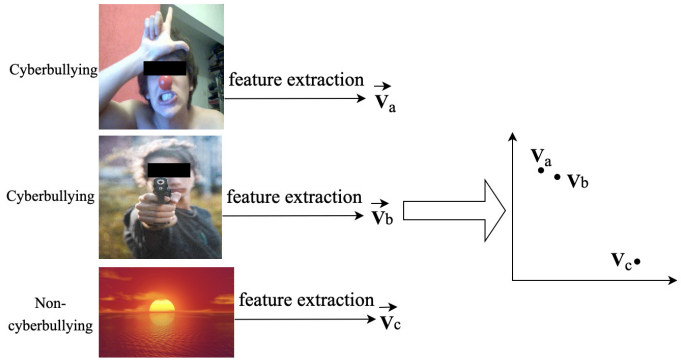


Fig. 10: Image feature representation: The first two images on the left panel have cyberbullying implication while the third image does not. Image feature representation intends to represent each image as a vector, showing on the right panel, with cyberbullying images being close to each other.

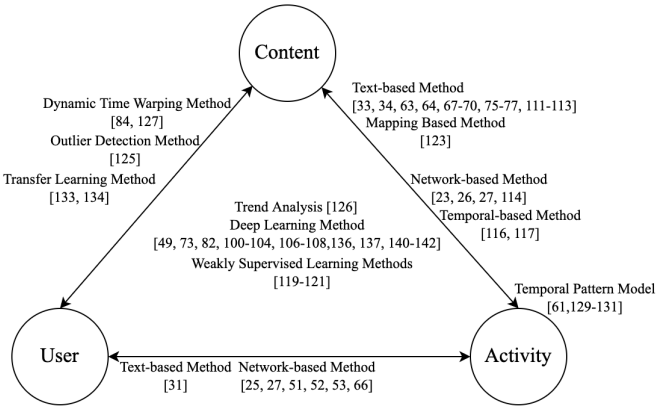


Fig. 11: A summary of cyberbullying detection methods and their focus on the UAC features

VIII. Conclusion

In this paper, we conducted a comprehensive review of cyberbullying and cyberviolence, with a focus on identifying key factors associated to cyberbullying and understanding interplay between these factors. We first proposed to summarize key factors of cyberbullying as a user-activity-content (UAC) triangular view. A feature taxonomy with three main categories, user-centered features, activity-centered features, and content-centered features, including sub-features within each category is proposed. Compared to existing work in the field, the UAC triangular view not only dissects seemly complicated features in cyberbullying detection into three major components: user, activity, and content, it also sets forth a clear understanding about how these features interplay with each other. Popular methods for cyberbullying detection are also reviewed, including supervised learning, weakly supervised learning, unsupervised learning, time series method, transfer learning and deep learning. The survey provides a thorough understanding of key factors, their strength and weakness, and available solutions for cyberbullying detection. The survey can also facilitate new designs of computational models for cyberbullying and cyberviolence detection.

TABLE XI: Method flexibility regarding with labels, features and adaptability

Method	Labels	Features	Adaptability
Supervised learning	+++	+++	+
Weakly supervised learning	++	+++	++
Unsupervised learning	-	+	++
Time series	+++	++	+
Transfer learning	++	++	+++
Deep learning	+	+/-	++

This survey provides many opportunities for future study on cyberbullying detection. First, our proposed UAC triangle organizes features commonly used in this area into three main categories (User, Activity and Content). Cyberbullying continuously evolves with the Internet. New features, such as location sharing, and regulations, like EU General Data Protection Regulation (GDPR), are being continuous made available. It is necessary to consider new feature types in the algorithm designs and system development. Second, due to page limitations, this review largely overlooked performance metrics, research projects, and actions taken by the industry and commercial systems to prevent cyberbullying. Third, literature from non-English sources can also be included to expand and enrich the scope of the review.

Acknowledgement

This research is partially sponsored by the U.S. National Science Foundation through Grant Nos. CNS-1828181, IIS-1763452, and by a seed grant of the College of Engineering and Computer Science, Florida Atlantic University.

References

[1] D. Olweus, "Victimization by peers: Antecedents and long-term outcomes," *Social withdrawal, inhibition, and shyness in childhood,NJ: Lawrence Erlbaum Associates*, pp. 315–341, 1993.

[2] P. K. Smith, K. C. Madsen, and J. C. Moody, "What causes the age decline in reports of being bullied at school? towards a developmental analysis of risks of being bullied," *Educational Research*, vol. 41, no. 3, pp. 267–285, 1999.

[3] N. C. P. Council. what is cyberbullying. [Online]. Available: https://www.ncpc.org/resources/cyberbullying/what-is-cyberbullying/

[4] A. Lenhart, M. Madden, A. Smith, K. Purcell, and K. Zickuhr. Teens, kindness and cruelty on social network sites. [Online]. Available: https://www.pewresearch.org/internet/2011/11/09/teens-kindness-and-cruelty-on-social-network-sites/

[5] E. Englander, E. Donnerstein, R. Kowalski, C. A. Lin, and K. Parti, "Defining cyberbullying," *Pediatrics*, vol. 140, no. Supplement 2, pp. S148–S151, 2017.

[6] T. Beran and Q. LI, "Cyber-harassment: A study of a new method for an old behavior," *Journal of Educational Computing Research - J EDUC COMPUT RES*, vol. 32, pp. 265–277, 2005.

[7] P. Bocij, *Cyberstalking: Harassment in the Internet age and how to protect your family*, first edition ed. Greenwood Publishing Group, 2004.

[8] B. W. B. H. Reyns and B. S. Fisher, "Being pursued online: Applying cyberlifestyle–routine activities theory to cyberstalking victimization," *Criminal justice and behavior*, vol. 38, no. 11, pp. 1149–1169, 2011.

[9] J. Allen and G. Norris, "Is genocide different? dealing with hate speech in a post-genocide society," *Journal of International Law & International Relations*, vol. 7, 2011.

[10] B. Perry, B. Levin, P. Iganski, R. Blazak, and F. Lawrence, *Hate crimes*. Westport, Conn. : Praeger Publishers, 2009.

TABLE XII: Summary of cyberbullying detection methods

Method	Strength	Weakness
Supervised learning	Good performance	Need large amount of labeled data
Weakly supervised learning	Less demanding of labeled data	Unstable performance
Unsupervised learning	No need labeled data	Low performance
Time series method	Capture temporal trend	Useful for short-term forecasting, but could lead to wrong predictions
Transfer learning	No need lots of new data	Negative transfer
Deep learning	No need feature engineering	Require very large amount of data and expensive to train

[11] J. Hawdon, A. Oksanen, and P. Räsänen, "Online extremism and online hate: Exposure among adolescents and young adults in four nations," *Nordicom Information*, vol. 37, pp. 29–37, 2015.

[12] N. C. for Education Statistics. Bullying at school and electronic bullying. [Online]. Available: https://nces.ed.gov/programs/coe/indicator/a10/

[13] J. Wang, R. Iannotti, and T. Nansel, "School bullying among adolescents in the united states: Physical, verbal, relational, and cyber," *The Journal of adolescent health*, vol. 45, pp. 368–75, 2009.

[14] L. Cheng, Y. N. Silva, D. Hall, and H. Liu, "Session-based cyberbullying detection: Problems and challenges," *IEEE Internet Computing*, vol. 25, no. 2, pp. 66–72, 2021.

[15] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbullying detection on twitter using big five and dark triad features," *Personality and Individual Differences*, vol. 141, pp. 252–257, 2019.

[16] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, pp. 1–14, 2020.

[17] B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying detection: A survey on multilingual techniques," in *2016 European Modelling Sym. (EMS)*. IEEE, 2016, pp. 165–171.

[18] A. Alakrot and N. S. Nikolov, "A survey of text mining approaches to cyberbullying detection in online communication flows," *NUI Galway-UL Alliance 5th Postgraduate Research Day*, 2015.

[19] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 3–24, 2020.

[20] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, "When the timeline meets the pipeline: A survey on automated cyberbullying detection," *IEEE Access*, vol. 9, pp. 103 541–103 563, 2021.

[21] K. Douglas, C. McGarty, A.-M. Bliuc, and G. Lala, "Groups understanding cyberhate: Social competition and social creativity in online white supremacist," *Social Science Computer Review - SOC SCI COMPUT REV*, vol. 23, pp. 68–76, 2005.

[22] K. Douglas, "Psychology, discrimination and hate groups on-line," *Oxford Handbook of Internet Psychology*, pp. 155–164, 2007.

[23] D. Chatzakou, I.Leoniadis, J.Blackburn, E.Cristofard, G.Stringhini, A.Vakali, and N.Kourtellis, "Detecting cyberbullying and cyberaggression in social media," *ACM Transactions on the Web*, vol. 13, no. 3, pp. 1–51, 2019.

[24] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," *Proc. of ACM WSDM 2010*, pp. 251–260, 2010.

[25] H. Hosseinmardi, S. Arredondo Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," *arXiv preprint arXiv:1503.03909*, 2015.

[26] Q. Huang, V. K. Singh, and P. K. Atrey., "Cyberbullying detection using social and textual analysis," *Proc. of the 3rd Intl. Workshop on Socially-Aware Multimedia, ACM*, pp. 3–6, 2014.

[27] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," *Proc. of ACM ASONAM*, p. 280–285, 2015.

[28] M. A. AL-GARADI, M. R. HUSSAIN *et al.*, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," *IEEE Access*, pp. 70 701–70 718, 2019.

[29] E. Whittaker and R. M. Kowalski, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, second edition ed. Elsevier, 2015.

[30] C. Duarte, S. K. Pittman, M. M. Thorsen, R. M. Cunningham, and M. L. Ranney, "Correlation of minority status, cyberbullying, and mental health: A cross-sectional study of 1031 adolescents," *Journal of Child & Adolescent Trauma*, vol. 11, no. 1, pp. 39–48, 2018.

[31] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," *in Proc. 25th DutchBelgian Inf. Retr. Workshop,*, pp. 1–3, 2012.

[32] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, , and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth." *Psychol Bull*, vol. 140, no. 4, p. 1073–1137., 2014.

[33] V. Nahar, S. AL Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," pp. 160–171, 2014.

[34] M. Dadvar, R. Ordelman, F. de Jong, and D. Trieschnigg, "Towards user modelling in the combat against cyberbullying," *Natural Language Processing and Information Systems*, pp. 277–283, 2012.

[35] M. Mccord and M. Chuah, "Spam detection on twitter using traditional classifiers," in *international conference on Autonomic and trusted computing*. Springer, 2011, pp. 175–186.

[36] P. T. Costa and R. R. McCrae, "Four ways five factors are basic," *Personality and Individual Differences*, vol. 13, no. 6, pp. 653–665, 1992.

[37] O. John and S. Srivastava, "The big five trait taxonomy: history, measurement, and theoretical perspectives," *Handb. Personal. Theory Res.*, vol. 2, pp. 102–138, 1999.

[38] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough, "The international personality item pool and the future of public-domain personality measures," *Journal of Research in Personality*, vol. 40, no. 1, pp. 84–96, 2006.

[39] L. R. Goldberg *et al.*, "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models," *Personality psychology in Europe*, vol. 7, no. 1, pp. 7–28, 1999.

[40] D. L. Paulhus and K. M. Williams, "The dark triad of personality: Narcissism, machiavellianism, and psychopathy," *Journal of Research in Personality*, vol. 36, no. 6, pp. 556–563, 2002.

[41] P. K. Jonason and G. D. Webster, "The dirty dozen: a concise measure of the dark triad." *Psychological assessment*, vol. 22, no. 2, p. 420, 2010.

[42] S. Jakobwitz and V. Egan, "The dark triad and normal personality traits," *Personality and Individual Differences*, vol. 40, no. 2, pp. 331–339, 2006.

[43] H. Douglas, M. Bore, and D. Munro, "Distinguishing the dark triad: Evidence from the five-factor model and the hogan development survey," *Psychology*, vol. 3, pp. 237–242, 2012.

[44] M. Geel, A. Goemans, F. Toprak, and P. Vedder, "Which personality traits are related to traditional bullying and cyberbullying? a study with the big five, dark triad and sadism," *Personality and Individual Differences*, vol. 106, pp. 231–235, 2017.

[45] R. Festl and T. Quandt, "Social relations and cyberbullying: the influence of individual and structural attributes on victimization and perpetration via the internet," *Human Communication Research*, vol. 39, pp. 101–126, 2013.

[46] A. K. Goodboy and M. M. Martin, "The personality profile of a cyberbully: Examining the dark triad," *Computers in Human Behavior*, vol. 49, pp. 1–4, 2015.

[47] R. Ang, K.-A. Tan, and M. Abu Talib, "Normative beliefs about aggression as a mediator of narcissistic exploitativeness and cyberbullying," *Journal of interpersonal violence*, vol. 26, pp. 2619–34, 12 2010.

[48] S. Pabian, C. J. De Backer, and H. Vandebosch, "Dark triad personality traits and adolescent cyber-aggression," *Personality and Individual Differences*, vol. 75, pp. 41–46, 2015.

[49] M. Yao, C. Chelmis, and D. Zois, "Cyberbullying ends here: Towards robust detection of cyberbullying in social media," *Proc. of WWW*, pp. 3427–3433, 2019.

[50] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee., "Measurement and analysis of online social networks," *Proc. of SIGCOMM ACM*, pp. 29–42, 2007.

[51] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: User classification in twitter," *SIGKDD Conf. on Knowledge Discovery and Data Mining*, p. 430–438., 2011.

[52] M. D. Choudhury, W. A. Mason, J. M. Hofman, and D. J. Watts., "Inferring relevant social networks from interpersonal communication," *Proc. of WWW ACM*, pp. 301–310, 2010.

[53] Mancilla-Caceres, J. F, D. Espelage, and E. Amir, "A computer game-based method for studying bullying and cyberbullying," *Journal of school violence*, vol. 14, no. 1, pp. 66–86, 2015.

[54] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment analysis for effective detection of cyber bullying," *Asia-Pacific Web Conference*, pp. 767–774, 2012.

[55] C. Ziems, Y. Vigfusson, and F. Morstatter, "Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification," *Proc. of ICWSM 2020*, 2020.

[56] E. Papegnies, V. Labatut, R. Dufour, and G. Linares, "Graph-based features for automatic online abuse detection," in *Intl. Conf. on statistical language and speech processing*. Springer, 2017, pp. 70–81.

[57] C. Chelmis, D. Zois, and M. Yao, "Mining patterns of cyberbullying on twitter," *Proc. of ICDM Workshop*, pp. 126–133, 2017.

[58] P. Bindu, P. S. Thilagam, and D. Ahuja, "Discovering suspicious behavior in multilayer social networks," *Computers in Human Behavior*, vol. 73, pp. 568–582, 2017.

[59] R. Kowalski and S. Limber, "Psychological, physical, and academic correlates of cyberbullying and traditional bullying," *The Journal of adolescent health*, vol. 53, pp. S13–20, 2013.

[60] D. Soni and V. Singh, "Time reveals all wounds: Modeling temporal dynamics of cyberbullying sessions," *12th ICWSM 2018*, pp. 684–687, 2018.

[61] A. Gupta, W. Yang, D. Sivakumar, Y. N. Silva, D. L. Hall, and M. N. Barioni, "Temporal properties of cyberbullying on instagram," *Companion Proceedings of the Web Conference 2020*, 2020.

[62] J. Kleinberg, "Bursty and hierarchical structure in streams," *Proc. of ACM SIGKDD Conference*, vol. 7, 2002.

[63] S. Ge, L. Cheng, and H. Liu, "Improving cyberbullying detection with user interaction," in *Proc. of the Web Conference 2021*, 2021, pp. 496–506.

[64] Noviantho, M. I. Sani, and A. Livia, "Cyberbullying classification using text mining," *IEEE ICICoS*, pp. 241–246, 11 2017.

[65] S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee, "A review of cyberbullying detection: An overview," in *13th ISDA*. IEEE, 2013, pp. 325–330.

[66] N. Tahmasbi and E. Rastegari, "A socio-contextual approach in automated detection of public cyberbullying on twitter," *ACM Trans. Soc. Comput.*, vol. 1, no. 4, 2018.

[67] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," *ICDCN, ACM 978-1-4503-4032-8*, pp. 1–6, 2016.

[68] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 3, pp. 1–30, 2012.

[69] N. Pendar, "Toward spotting the pedophile telling victim from predator in text chats." *Proc. of the Intl. Conf. on Semantic Computing,*, p. 235–241, 2017.

[70] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," *Proc. of NAACL-HLT2012*, pp. 656–666, 2012.

[71] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, *A Practical Guide to Sentiment Analysis*. Springer Intl. Publishing, 2017, vol. 5.

[72] S. Murnion, W. J. Buchanan, A. Smales, and G. Russell, "Machine learning and semantic analysis of in-game chat for cyberbullying," *Computers & Security*, vol. 76, pp. 197–213, 2018.

[73] M. A. Al-Ajlan and M. Ykhlef, "Optimized twitter cyberbullying detection based on deep learning," in *2018 21st Saudi Computer Society National Computer Conference (NCC)*. IEEE, 2018, pp. 1–5.

[74] J. L. Bigelow, A. Edwards, and L. Edwards, "Detecting cyberbullying using latent semantic indexing," in *Proc. of the 1st Intl. workshop on computational methods for CyberSafety*, 2016, pp. 11–14.

[75] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," *Association for the Advancement of Artificial Intelligence (www.aaai.org)*, pp. 11–17, 2011.

[76] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang., "Abusive language detection in online user content." *Proc. of TheWebConf*, p. 145–153., 2016.

[77] D. Yin, Z. Xue, L. Hong, B. Davison, A. Edwards, and L. Edwards, "Detection of harassment on web 2.0," *CAW2.0,2009, Madrid, Spain*, 2009.

[78] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *Proc. of the 31st Intel. Conf. on Machine Learning, Beijing, China*, pp. 1188–1196, 2014.

[79] N. Aulia and I. Budi, "Hate speech detection on indonesian long text documents using machine learning approach," *ICCAI '19,Bali, Indonesia, ACM,*, pp. 164–169, 2019.

[80] P. Burnap and M. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making: Machine classification of cyber hate speech," *Policy & Internet*, vol. 7, 2015.

[81] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A multilingual evaluation for online hate speech detection," *ACM Transactions on Internet Technology (TOIT)*, vol. 20, pp. 1 – 22, 2020.

[82] H. H. Nishant Vishwamitra and, F. Luo, and L. Cheng†, "Towards understanding and detecting cyberbullying in real-world images," *Network and Distributed Systems Security (NDSS) Sym.*, 2021.

[83] F. Patacsil, "Analysis of cyberbullying incidence among filipina victims: A pattern recognition using association rule extraction," *Intelligent Systems and Applications*, pp. 48–57, 2019.

[84] N. Potha and M. Maragoudakis, "Cyberbullying detection using time series modeling," *ICDMW*, vol. 2014, pp. 373–382, 2014.

[85] D. Robinson, Z. Zhang, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network." *Proc. of (ESWC'18)*, pp. 745–760., 2018.

[86] S. O. Sood, E. F. Churchill, and J. Antin, "Automatic identification of personal insults on social news sites," *J. of the American Society for Information Science and Tech.*, vol. 63, no. 2, pp. 270–285, 2012.

[87] K. Wang, Y. Cui, J. Hu, Y. Zhang, W. Zhao, and L. Feng, "Cyberbullying detection, based on the fasttext and word similarity schemes," *ACM TALLIP*, vol. 20, pp. 1 – 15, 2021.

[88] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proc. of NAACL-HLT*, 2012, pp. 656–666.

[89] M. Dadvar and F. de Jong, "Cyberbullying detection; a step toward a safer internet yard," *ACM 978-1-4503-1230-1/12/04.*, 2012.

[90] K. Burn-Thornton and T. Burman, "The use of data mining to indicate virtual (email) bullying," *Pro. of GCIS 2012*, pp. 253–256, 2012.

[91] S. O. Sood, J. Antin, and E. Churchill, "Using crowdsourcing to improve profanity detection," in *2012 AAAI Spring Sym. Series*, 2012.

[92] U. Bretschneider, T. Wöhner, and R. Peters, "Detecting online harassment in social networks," *Thirty Fifth International Conference on Information Systems, Auckland 2014*, pp. 1–14, 2014.

[93] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Štefan Emrich, and M. Sedlmair., "More than bags of words: Sentiment analysis with word embeddings." *Communication Methods and Measures 12, 2-3 (2018),*, p. 140–157., 2018.

[94] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," *European Conference on Information Retrieval,Springer,*, pp. 693–696., 2013.

[95] M. Fortunatus, P. Anthony, and S. Charters, "Combining textual features to detect cyberbullying in social media posts," *Procedia Computer Science 176*, p. 612–621, 2020.

[96] M. Dadvar, D. Trieschnigg, and F. de Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies," *Canadian Conf. on Artificial Intelligence*, pp. 275–281, 2014.

[97] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13 825 – 13 835, 2018.

[98] L. Hamers *et al.*, "Similarity measures in scientometric research: The jaccard index versus salton's cosine formula." *Information Processing and Management*, p. 315–318, 1989.

[99] P. Blandfort, D. U. Patton, W. R. Frey, and others., "Multimodal social media analysis for gang violence prevention." *Proc. of the AAAI Conf. on Web and Social Media*, vol. 13, p. 114–124, 2019.

[100] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," *IEEE/ACM ASONAM*, pp. 186–192, 2016.

[101] V. K.Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," *Proc. of ACM CHI*, p. 2090–2099, 2017.

[102] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, "Content-driven detection of cyberbullying on the instagram social network." *IJCAI*, pp. 3952–3958, 2016.

[103] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications." *Proc. pf IEEE/CVF WACV*, 2020.

[104] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *In International conference on machine learning*, p. 2048–2057, 2015.

[105] M. Khan, M. Tahir, and Z. Ahmed, "Detection of violent content in cartoon videos using multimedia content detection techniques," *2018 IEEE 21st INMIC*, pp. 1–5, 2018.

[106] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova., "Privacy-aware image classification and search,," *Proc. of ACM SIGIR*, p. 35–44, 2012.

[107] Y. Miyakoshi and S. Kato, "Facial emotion detection considering partial occlusion of face using bayesian network," *2011 IEEE Sym. on Computers & Informatics*, pp. 96–101, 2011.

[108] A. Das, J. S. Wahi, and S. Li, "Detecting hate speech in multi-modal memes," *arXiv preprint arXiv:2012.14891*, 2020.

[109] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[110] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, 2007.

[111] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *IEEE Trans. on Big Data*, vol. 6, no. 1, pp. 3–28, 2020.

[112] N. Tahmasbi and A. Fuchsberger, "Challenges and future directions of automated cyberbullying detection," in *AMCIS 2018*. Association for Information Systems, 2018.

[113] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[114] A. Wang and K. Potika, "Cyberbullying classification based on social network analysis," in *2021 IEEE BigDataServic*. IEEE Computer Society, 2021, pp. 87–95.

[115] P. K. A. Vivek K. Singh, Qianjia Huang, "Cyberbullying detection using probabilistic socio-textual information fusion." *IEEE/ACM,ASONAM ,978-1-5090-2846-7*, pp. 884–887, 2016.

[116] T. G. Dietterich, "Ensemble methods in machine learning." *Proc. of the 1st Intl. Workshop on Multiple Classifier Systems.*, 2000.

[117] M. F. López-Vizcaíno, F. J. Nóvoa, V. Carneiro, and F. Cacheda, "Early detection of cyberbullying on social media networks," *Future Generation Computer Systems*, vol. 118, pp. 219–229, 2021.

[118] F. Cacheda, D. Fernandez, F. J. Novoa, V. Carneiro *et al.*, "Early detection of depression: social network analysis and random forest techniques," *Journal of medical Internet research*, vol. 21, no. 6, p. e12554, 2019.

[119] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, pp. 44–53., 2018.

[120] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Australasian Database Conference*. Springer, 2014, pp. 160–171.

[121] E. Raisi and B. Huang, "Cyberbullying detection with weakly supervised machine learning," *Proc. of ASONAM ACM*, pp. 409–416, 2017.

[122] E. Raisi and B. Huang., "Co-trained ensemble models for weakly supervised cyberbullying detection." *In NIPS LLD Workshop.*, 2017.

[123] T. Kohonen, "Self-organizing maps: Ophmization approaches," in *Artificial Neural Networks*. North-Holland, 1991, pp. 981–990.

[124] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 432–437, 2016.

[125] A. Rauber, D. Merkl, and M. Dittenbach, "The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data," *IEEE Trans. on Neural Networks*, vol. 13, no. 6, pp. 1331–1341, 2002.

[126] U. Bretschneider, T. Wöhner, and R. Peters, "Detecting online harassment in social networks," *Thirty Fifth International Conference on Information Systems*, pp. 1–14, 2014.

[127] S. Das, A. Kim, and S. Karmakar, "Change-point analysis of cyberbullying-related twitter discussions during covid-19," *16th Annual Social Informatics Research Sym.*, 2020.

[128] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[129] E. Keogh and M. Pazzani, "Derivative dynamic time warping," *First SIAM Intl. Conf. on Data Mining*, vol. 1, 2002.

[130] L. Cheng, R. Guo, Y. N. Silva, D. Hall, and H. Liu, "Modeling temporal patterns of cyberbullying detection with hierarchical attention networks," *ACM/IMS Trans. on Data Science*, vol. 2, no. 2, pp. 1–23, 2021.

[131] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," in *Proc. SIAM Intl. Conf. on data mining*. SIAM, 2019, pp. 235–243.

[132] D. Soni and V. Singh, "Time reveals all wounds: Modeling temporal characteristics of cyberbullying," in *Proc. of the AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018.

[133] S. J. Pan and Q. Yang, "A survey on transfer learning." *Trans. on Knowledge and Data Eng. 22(10), IEEE*, p. 1345–1359, 2010.

[134] M.-A. Rizoiu, T. Wang, G. Ferraro, and H. Suominen, "Transfer learning for hate speech detection in social media," *arXiv preprint arXiv:1906.03829*, 2019.

[135] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," *European conference on information retrieval, Springer*, pp. 141–153, 2018.

[136] J. Pennington, R. Socher, and C. D. M. C, "Glove: Global vectors for word representation." *In EMNLP*, p. 1532–1543, 2014.

[137] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa., "Natural language processing (almost) from scratch." *Journal of Machine Learning Research,12*, p. 2493–2537., 2011.

[138] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification." *Proc. of ACL-IJCNLP.*, p. 1555–1565, 2014.

[139] T. N. Kipf and M. Welling, "Variational graph auto-encoders." *arXiv preprint arXiv:1611.07308*.

[140] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. of 22nd ACM SIGKDD*, 2016, pp. 855–864.

[141] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[142] K. Kumari and J. P. Singh, "Identification of cyberbullying on multi-modal social media posts using genetic algorithm," *Tran. on Emerging Telecommunications Technologies*, vol. 32, no. 2, p. e3907, 2021.

[143] P. Linforth. Teens, kindness and cruelty on social network sites. [Online]. Available: https://pixabay.com/illustrations/sunset-sea-horizon-sun-sky-ocean-298850/

Shuwen Wang is a PhD student in the Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA. She joined FAU PhD program in 2020, and her research mainly focuses on data mining, machine learning, and medical data analysis.

Xingquan Zhu (SM'12) is a Full Professor in the Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA. His research interests include data mining, machine learning, and bioinformatics. Since 2000, he has published more than 280 refereed journal and conference papers in these areas, including four Best Paper Awards (PAKDD-21, IRI-18, PAKDD-13, ICTAI-05) and three Best Student Paper Awards (ICDM-20, ICKG-20, ICPR-12). He is the Program Committee Co-Chair for the 22nd IEEE International Conferene on Data Mining (ICDM-2022), General Co-chair for the 2021 IEEE International Conference on Big Data (IEEE BigData-2021), and Program Committee Co-Chair for the 33rd International Conference on Scientific and Statistical Database Management (SSDBM-2021). He previously served as an associate editor of the IEEE Trans. on Knowledge and Data Engineering (2008-2012, 2014-2021), and currently serves an associate editor of the ACM Trans. on Knowledge Discovery from Data (2017 - date).

Weiping Ding (M'16-SM'19) is a professor and the Vice Dean of the School of Information Science and Technology, Nantong University, Nantong, China, and also the supervisor of Ph.D postgraduate by the Faculty of Data Science at City University of Macau, Macau, China. His research interests include deep neural networks, multimodal machine learning, granular data mining, and medical images analysis. He has published over 120 scientific articles in refereed international journals, such as IEEE T-FS, T-NNLS, T-CYB, T-SMCS, T-BME, T-EVC, T-II, T-ETCI, T-CDS, T-ITS and T-AI. He has held 20 approved invention patents. He has co-authored two books. His nine authored/co-authored papers have been selected as ESI Highly Cited Papers. Dr. Ding served/serves on the Editorial Board of Knowledge-Based Systems, Information Fusion, Engineering Applications of Artificial Intelligence and Applied Soft Computing. He served/serves as an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Fuzzy Systems, IEEE/CAA Journal of Automatica Sinica, Information Sciences, Neurocomputing, Swarm and Evolutionary Computation, and so on. He is the Leading Guest Editor of Special Issues in several prestigious journals, including IEEE Transactions on Evolutionary Computation, IEEE Transactions on Fuzzy Systems..

Amir Alipour Yengejeh is a PhD student in the Dept. of Electrical Engineering & Computer Science, Florida Atlantic University, Boca Raton, FL, USA. His research mainly focuses on data mining, machine learning, and statistical data analysis.