Analysing and Finding of Cyberbullyies In Social Media Based on Data Science And Web Application

Madhavan S[1], Praveen M[2], Mrs.A.S.Hepsi Ajibah[3]
[1, 2, 3] Dept of Information Technology
[3]Professor, Dept of Information Technology
[1, 2, 3] GKMCET Chennai, Tamilnadu, India.

Abstract- *One of the harmful consequences of social media is the rise of cyber bullying, which tends to be more sinister then traditional bullying given that online records typically live on the internet for quite a long time and are hard to control. We exploit bullying tendencies by proposing a robots method for constructing a cyber bullying signed network. We analyze tweets to determine their relation to cyber bullying. While considering the context in which the tweets exits in order to optimize their bullying score. We are going to implement an algorithm to detect cyber bullies and their messages using data science.With the increased utilization of the internet and social media platforms, it is not surprising that youth are using these tools to inflict harm upon each other. The purpose of this paper is to explore the pervasiveness of cyber bullying among university students in an Arab community, its nature and venues, and their attitudes towards reporting cyber bullying in contrast to remaining silent. Data were collected from 200 students in the UAE. 91% of the study sample confirmed the existence of acts of cyber bullying on social media with Instagram (55.5%) and Facebook (38%) in the lead. Calls for smartphone applications, stricter legal actions and proactive measures are discussed.*

Keywords- Cyber bullying, Twitter classification,Online shaming, social media,Graph analysis, Proactive Measures.

I. INTRODUCTION

Bullying is defined as intentional aggression carried out repeatedly by one individual or a group of individuals towards a person who is unable to easily defend him or herself (Olweus, 1993). Cyber bullying is, by extension, defined by Smith et al. (2008, pg. 376) as "an aggressive, intentional act carried out by a group or individual using electronic forms of contact, repeatedly or over time against a victim that cannot easily defend him or herself". Hinduja and Patchin (2009) define cyber bullying as "wilful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices". Cyber bullying has been found to be quite prevalent on social media with as many as 54% of young people reportedly cyber bullied on Facebook (Ditch The Label, 2013). Zhang et al. (2016) found that neutralizing processes (Sykes and Matza, 1957) play a significant role in why many young people engage in cyber bullying. They surmised that cyber bullies engage in such delinquent acts by rationalizing their behaviors as valid and that the severity of possible sanctions does not deter.

There is substantial variation in the reported frequency for cyber bullyingvictimization, with rates as low as 45% reported by Olweus (2012) for the U.S.A. and rates as high as 35%–57% reported for mainland China (Zhou *et al*., 2013). Patching and Hinduja (2012) reported a frequency of about 20% amongst their survey of 4,400students and found an average rate of 24% across existing studies. The EU Kids Online report (Livingston *et al.*, 2014) surmised that cyber bullying has now surpassed face-to-face bullying in the UK, with 12% of teenagers aged 9–16 years experiencing some form of cyber bullyingvictimization as opposed to 9% for face-to-face bullying. This variation in the reported frequency of cyber bullying has been attributed to how cyber bullying has been defined by each study (Patchin and Hinduja, 2012) and the length of the intervening period between a cyber bullying incident and when victims were interviewed (Sabella*et al*., 2013), with (perhaps unsurprisingly) the more recent victims of cyber bullying scoring higher on impacts and effects.

The detection of cyber bullying and online harassment is often formulated as a classification problem. Techniques typically used for document classification, topic detection, and sentiment analysis can be used to detect electronic bullying using characteristics of messages, senders, and the recipients. It should, however, be noted that cyber bullying detection is intrinsically more difficult than just detecting abusive content. Additional context may be required to prove that an individual abusive message is part of a sequence of online harassment directed at a user(s) for such a message to be labeled as cyber bullying. Thus, a tweet such as "*@username So you got drunk at a party and two people take advantage of you, that's not rape you're just a loose drunk slut #BiasedResults #Steubenville*" can be easily classified as online harassment due its use of profanity ("slut") but requires additional context such as conversation history to determine if

this is indeed bullying. Cyber bullying detection is inherently difficult due to the subjective nature of bullying. It extends beyond detecting negative sentiments or abusive content in a message as these tasks, on their own, do not necessarily mean that the message is in fact bullying. For example, a message such as "*I'm disgusted by what you said today and I never want to see you again*" is difficult to classify as bullying without understanding the larger context of the exchange, even though the message is clearly expressing very negative sentiments. Conversely, positively-expressed sentiments may disguise bullying if the intent is to express sarcasm.

We use this definition as part of our survey's inclusion criteria and only include studies that attempt one or more of the above tasks. In defining the roles identification task, we used the 8 roles identified by Xu*et al.* (2012a) as the superset of roles. These are of *bully*, *victim*, *bystander*, *assistant*, *defender*, *reporter*, *accuser*, and *reinforcer*. *Bystanders* are witnesses that do not intervene in a bullying incident. *Assistants* are co-perpetrators but not initiators. *Reinforcers*, while not directly involved in the bullying, encourage bullies and provide an impetus for continuation (e.g., laughing at the expense of victims). An *accuser* differs from a *reporter* by actively identifying *victim*s and *bullies*. Finally, *defenders* aid victims by coming to their aid. These roles encompass the various roles actors can inhabit during a cyber bullying incident and, as such, our sample includes studies that detected one or more of these roles. In fact, we did not find any study that attempted detecting roles outside of these 8 roles.

II. DATA SEARCH AND SELECTION

An electronic literature search was conducted across Scopus, the ACM Digital Library, and the IEEE Explore digital library. The main search strategy was the discovery of academic literature relevant to the theme "automated detection of electronic bullying, anti-social behaviour and harassment" using the following query phrases without any publication year filter applied:

"cyber-bull* or cyber bull* detection", "detecting cyber bull* or cyber bull*", "electronic or online bullying detection", "detecting electronic or online bullying, cyber-bull*" or "cyber bull* prevention tool", "cyber-bull* or cyberbull* prevention software", "cyber-bull* or cyberbull* software", "anti cyber-bull* or anti cyberbull*" or "anti-cyberbull* or anti-cyber-bull*" or "anticyberbull* or anticyberbull*", "detecting electronic or online harassment".

A citation trail was performed on the discovered papers using the papers' references as a starting point and a total of 89 academic papers was discovered as a result of the search. The papers were initially assessed for relevance via a review of their titles, abstract, and concluding arguments: 18 papers were not considered relevant to the survey and so were removed. The full text of the remaining papers was reviewed and papers whose primary focus did not include any of the 4 cyberbullying detection tasks we identified in Section 1 were discounted. This led to the removal of a further 18 papers. These included papers that dealt with themes such as youth violence involvement detection (Sigel and Harpin, 2013), story matching to identify distressed teens (Dinakar*et al*., 2012b; Macbeth *et al*., 2013), and cyberbully prevention policies (Al Mazari, 2013). To eliminate the effects of language on cyberbully detection when comparing the reviewed studies, we excluded papers using non-English corpora; thus a further 7 papers were excluded. These included papers such as Ptazynski*et al*. (2010a; b), Honjo*et al.* (2011), Nitta *et al.* (2013), Li and Tagami (2014), Margono*et al.* (2014) and Van Hee*et al.* (2015) which were removed as they used non-English corpora.

III. FEATURES USED FOR CYBERBULLYING DETECTION

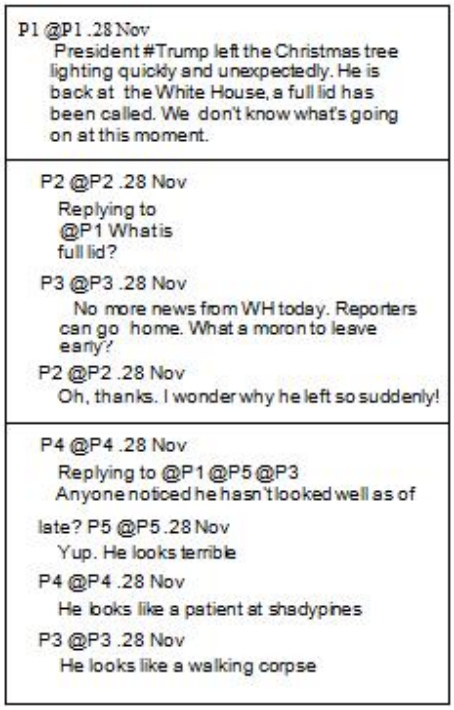
We broadly categorise features used across the studies into 4 main groups, namely content-, sentiment-, user- and network-based features. We define content-based features as the extractable lexical items of a document such as keywords, profanity, pronouns, and punctuations. Emotion based features are those features that are indicative of emotive content; they are generally keywords, phrases and symbols (e.g., emoticons) that can be used to determine the sentiments expressed in a document. User-based features are those characteristics of a user's profile that can be used to make a judgment on the role played by the user in an electronic exchange and include age, gender, and sexual orientation. Finally, network-based features are usage metrics that can be extracted from the online social network and include items such as number of friends, number of followers, frequency of posting, etc.

IV. CYBERBULLYING DETECTION

Much work has been done over the past decade in the area of cyberbullying detection. There have been two broad approaches in identifying bullies - one aims to detect bullying messages [65, 50, 23, 63, 14, 15, 16], while the other approach is to detect the cyberbullies responsible for the messages [51, 18, 10, 11].

The first approach is to determine bullying messages, some used text-based analytics, and others used a mix of text

and user features. Zhao *et al*. [65] proposed a text based Embeddings-Enhanced Bag-of-Words (EBoW) model that utilizes a concatenation of bul lying features, bag-of-words, and latent semantic features to obtain a final representation, which is then passed through a classifier to identify cyberbullies. Xu*et al*. [63] used textual information to identify emotions in bullying traces, as opposed to determining whether or not a message was bullying. Singh *et al*. [50] proposed a probabilistic socio-textual information fusion for cyberbullying detection. This fusion uses social network features derived from a 1.5 ego network and textual features, such as density of bad words andpart-of-speech-tags. Hosseinmardi*et al.*, used images and text to detect cyberbullying.



V. PROPOSED METHOD

objective of our solution is to identify bullies from raw Twitter data based on the context as well as the contents in which the tweet exists. Given a set of tweets *T* containing the Twitter features such as user ID, reply ID etc, our approach consists of three algorithms - (i) Conversation Graph Generation Algorithm, (ii) Bullying Signed Network Generation Algorithm and (iii) Bully Finding Algorithm. The first algorithm constructs a directed weighted conversation graph $G_c$by efficiently reconstructing the conversations from raw Twitter data while enabling a more accurate model of human interactions. The second algorithm constructs a bullying signed network B to analyze the behaviour of users in a social media. Finally, the third algorithm consists of our proposed attitude and merit centrality measures to identify bullies from B. Figure 5.1 shows the process flow of BullyNetwhere the raw data is extracted from Twitter using Twitter API from which the conversation graph is constructed for each conversation using algorithm 5.1. Then from the conversation graphs, a bullying signed network is generated using algorithm 5.2. Finally the bullies from Twitter are identified by applying algorithm

A. Conversation Graph Generation

The conversation graph generation algorithm 5.1, is constructed from a set of tweets $T = \{t_1, \ldots, t_n\}$ to generate a directed weighted conversation graphs $G_c = \{gc^{1}, \ldots, gcm\}$ for each conversations c_i, which is extracted from the tweets T . The graphs are represented as $G_c = (V, E)$ where V is the set of users involved in the conversation, E is the set of edges representing the tweets in the conversation, and each edge is assigned a bullying indicator value I as the edge weight which is in the range of $[-1, +1]$. When $I_{ij} = -1$, it indicates the negative interaction by $i$towards j and when $I_{ij} = 1$, it indicates the positive interaction by $i$towards j. The bullying indicator I, for each tweet is calculated based on sentiment analysis and cosine similarities. In Step 1, the tweets set T are sorted based on the creation time to reduce the time complexity, while searching for tweets based on DID. Moreover, the set is sorted in a reverse-chronological order so that every DID of a tweet matches with only one SID of the remaining tweets. In Step 2, for each tweet $t_i$in T , the conversations are built by doing a binary search DID(t_i) with the SID of the remaining tweets. If a match is found ast^{J}then, it is appended with $t_i$to form a new conversation. If binary search match is found with the already existing tweet in the conversation c_i then, $t_i$is appended to tweets in c_i.

B. Bullying Signed Network Generation

In many real-world social systems, the relation between two nodes can be represented as signed networks with positive and negative links. Since this research focuses on identifying the bullying nodes in the network, the algorithm 5.2 is designed to determine the final outgoing edge weight, wijfor the users in the conversation graphs Gc. In Step 1a, for every conversation graph gci, a bullying score S is calculated for the users(nodes) in that graph based on the tweet order (sorted in ascending order). For an edge e = (u, v), the bullying score Suvis set to Iuvif the edge towards v is not a reply from u or else, the bullying score Suvis calculated as Iuv + ((Iuv− Svu) ∗α) where α is a constant which will be determined by the experiment. If there are more than one edge for a user with the same order then, after the bullying

score is evaluated, an average bullying score is computed for the same set of order.

C. Bully Finding

Given a BSN, with a graph $G_s = (V, E, W)$, where V is the set of users as nodes and E is the set of edges directed from node $i$to node j, has weight $w_{ij} \in W$ withinthe range [-1,1]. Our research is to identify bullies from B using centrality measure. G E Centrality is a measure in a network that is used to identify the most important vertices and 2 also to determine how one vertex affects others in a network. The importance of a vertex 0 or node is determined by how high the score is within a network and also defined by the type of the network. Since this research is about social networks the importance is defined as the behaviour. Among several centrality measures, we consider Bias and Deserve (BAD) by Mishra and Bhattacharya [41] because, their measure is computed on how the outgoing edge from a node/user depends on the incoming edges from other nodes/users. However, BAD is modelled on a trust based network i.e., the users that have a propensity to trust/distrust other users. Also, the edge weight denotes trust score rather than the bullying score as in this research.

VI. IMPLEMENTATION AND SETUP

We implemented our algorithm in Java, and our experiments were conducted on a machine equipped with an Intel(R) Core(TM) i7-8550U CPU @ 2.00GHz processor and 16.0 GB RAM, running Windows 10 64-bit operating system. We employed Amazon *Mechanical Turk* (mturk) workers to respond to an online survey that we developed. We provided 2700 surveys with each survey consisting of 10 2conversations. Each survey was assigned to three workers to classify the bullying 0 behavior of the users in the conversations according to predefined labels (strongly positive, likely positive, likely negative and strongly negative). Overall, the workers rated 27000 conversations, which were extracted from the set of raw Twitter data by using algorithm 5.1. The MTurk UI enables requesters to create and publish HITs in a batch when processing many HITs of the same type thus saving time. For our study, we created a csv file that contained 2700 HITs. MTurk automatically created a separate HIT for each set of conversation in the csv file. The results to rate each users involved in the set of conversations are obtained from the workers. There was not marked variation in rating provided by the workers. Finally, the results are combined for the users to form the ground truth.

VII. CONCLUSION

Although the digital revolution and the rise of social media enabled great advances in communication platforms and social interactions, a wider proliferation of harmful behavior known as bullying has also emerged. Aiming to address this bullying, this thesis presents a novel framework to identify bully users from the Twitter social network. We performed extensive research on mining signed networks for better understanding of the relationships between users in social media, to build a signed network (SN) based on bullying tendencies. We observed that by constructing conversations based on the context as well as content, we can effectively identify the emotions and the behavior behind bullying. In our experimental study, the evaluation of our proposed centrality measures to detect bullies from signed network, we achieved 70% accuracy with 77% precision in identifying bullies.

we presented and discussed the building blocks of our thesis that are very important for the implementation of this work. We explain the sentiment analysis, the different techniques to analyze the sentiment of the message, the cosine similarity, the centrality measures and the different types of measures used in signed networks.we examined the related work done in the field of cyberbullying detection. We did extensive research on a signed network focusing on node classification, balance theory and measure designed to analyze the signed network. We made a comparativeevalution.

REFERENCES

[1] ApoorvAgarwal, FadiBiadsy, and Kathleen R Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the Conference of the European Chapter of the ACL*, pages 24–32, 2009.

[2] SmritiBhagat, Graham Cormode, and S Muthukrishnan. Node classification in social networks. In *Social network data analytics*, pages 115–148. 2011.

[3] PetkoBogdanov, Nicholas D Larusso, and Ambuj Singh. Towards community discovery in signed collaborative interaction networks. In *Proceedings of the IEEE International Conference on DMW*, pages 288–295, 2010.

[4] Phillip Bonacich and Paulette Lloyd.Calculating status with negative relations.*Social networks*, 26(4):331–338, 2004.

[5] PiotrBorzymek and MarcinSydow. Trust and distrust prediction in social network with combined graphical and review-based attributes. In *Proceedings of the KES AMSTA*, pages 122–131, 2010.

[6] UlrikBrandes and Dorothea Wagner.Analysis and visualization of social networks. In *Graph drawing software*, pages 321–340. 2004.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextualweb search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[8] Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus.*Trolls just want to have fun*, pages 67:97–102. 2014.

[9] Cyberbullying Research Center. https://cyberbullying.org/bullying-laws.

[10] DespoinaChatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, GianlucaStringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the ACM on WebSci*, pages 13–22, 2017.

[11] Ying Chen, Yilu Zhou, Sencun Zhu, and HengXu. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the PASSAT and SCSM*, pages 71–80, 2012.

[12] Kai-Yang Chiang, NagarajanNatarajan, AmbujTewari, and Inderjit S Dhillon. Ex G
ploiting longer cycles for link prediction in signed networks. In *Proceedings of the ACM international CIKM*, pages 1157–1162, 2011.

[13] Council Of Europe children's rights. https://www.coe.int/en/web/children/bullying.

[14] MaralDadvar, DolfTrieschnigg, Roeland Ordelman, and Franciska de Jong. Im proving cyberbullying detection with user context. In *Proceedings of the European Conference on IR*, pages 693–696, 2013.

[15] KarthikDinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyber bullying.*In Proceedings of the ACM TiiS*, 2(3):18, 2012.

[16] KarthikDinakar, RoiReichart, and Henry Lieberman. Modeling the detection of textualcyberbullying. In *Proceedings of the international AAAI WSM*, 2011.

[17] Thomas DuBois, Jennifer Golbeck, and AravindSrinivasan. Predicting trust and dis trust in social networks. In *Proceedings of the IEEE international PASSAT conference on SC*, pages 418–424, 2011.

[18] PatxiGaln-Garca, J.G. De La Puerta, C.L. Gmez, Igor Santos, and Pablo Bringas. Su pervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. 24:42–53, 2014.

[19] LiseGetoor and Christopher P Diehl. Link mining: a survey. *In Proceedings of the ACM SIGKDD Explorations Newsletter*, 7(2):3–12, 2005.

[20] RamanthanGuha, Ravi Kumar, PrabhakarRaghavan, and Andrew Tomkins.Propagation of trust and distrust. In *Proceedings of the international conference on WWW*, pages 403–412, 2004.

[21] Jiawei Han, Jian Pei, and MichelineKamber.*Data mining: concepts and techniques*. 2011.

[22] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the international conference on WWW*, pages 517–526, 2002.

[23] HomaHosseinmardi, Sabrina Arredondo Mattson, RahatIbnRafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. *In Proceedings of the CoRR*, abs/1503.03909, 2015.

[24] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceed ings of the ACM SIGKDD international conference on KDD*, pages 168–177, 2004.

[25] Xia Hu, Jiliang Tang, HuijiGao, and Huan Liu. Social spammer detection with sentiment information. *In Proceedings of IEEE ICDM*, pages 180–189, 2014.