Introduction

Domain knowledge

The Internet has created never before seen opportunities for human interaction and socialization. While the growth of social media has created an excellent platform for communications and information sharing, it has also created a new platform for malicious activities, such as spamming, trolling, and cyberbullying. According to the Cyberbullying Research Center (CRC), cyberbullying occurs when someone uses the technology to send messages to harass, mistreat, or threaten a person or a group. So a system is required to detect cyberbullying.

Motivation/Objectives

One of the most harmful consequences of social media is the rise of cyberbullying, which tends to be more sinister than traditional bullying, given that online records typically live on the Internet for quite a long time and are hard to control. So to avoid them the system is proposed with following objectives:

- To collect, preprocess, and label the Twitter data set.
- To Propose a novel efficient algorithm for detecting cyberbullies on Twitter.
    1. To build conversation.
    2. To construct bullying SN.
    3. Propose A&M centrality.

Literature Review

- Wan Noor Hamiza Wan Ali, et al., [1] discussed about cyberbullying detection, available data sources, features and classification techniques. Natural Language Processing (NLP) and machine learning are the famous approaches used to identify bullying keywords within the corpus.
- Djedjiga Mouheb, et al., [2] presented an approach to detect cyberbullying in Arabic Twitter streams in real-time. In addition, it classifies the bullying messages based on their strength. In case a cyberbullying message is detected, the system notifies the user and proposes a set of actions to take based on the strength of the bullying message.
- Vijay Banerjee, et al., [3] proposed a novel cyberbullying detection method dependent on deep neural network. Convolution Neural Network is utilized for the better outcomes when contrasted with the current systems.
- Vivek K. Singh, et al., [4] audited an existing cyberbullying algorithm using Twitter data for disparity in detection performance based on the network centrality of the potential victim and then demonstrate how this disparity can be countered using an Equalized Odds post-processing technique. The results pave the way for more accurate and fair cyberbullying detection algorithms.
- Ong Chee Hang, et al., [5] proposed a cyberbullying lexicon for social media. consisting of several phases, namely: understanding the concepts of exclusion cyberbullying, word list selection, keyword identification, classes and subclasses identification, and lastly cyberbullying ontology and lexicon development.
- Jason Wang, et al., [6] aimed to investigate the viability of an automatic multiclass cyberbullying detection model that is able to classify whether a cyberbully is targeting a victim's age, ethnicity, gender, religion, or other quality. They have established a framework for the automatic generation of balanced data by using a semi-supervised online Dynamic Query Expansion (DQE) process to extract more natural data points of a specific class from Twitter. And also proposed a Graph Convolutional Network (GCN) classifier, using a graph constructed from the thresholded cosine similarities between tweet embeddings. With DQE-augmented dataset, compared the GCN model using eight different tweet embedding methods and six other classification models over two sizes of datasets.

- Ankit Pradhan, et al., [7] explored the adaptivity and efficiency of self-attention models in detecting cyberbullying. experimented with the Wikipedia, Formspring and Twitter cyberbullying datasets and achieve more efficient results over existing cyberbullying detection models. Also proposed new research directions within cyberbullying detection over recent forms of media like Internet memes which pose a variety of new and hybrid problems.
- Antonio Calvo-Morata, et al., [8] developed Conectado, a serious game to be used in the classroom to increase awareness on bullying and cyberbullying in schools. While playing the game, students gain a first-hand immersive experience of the problem and the associated emotions, fostering awareness and empathy with victims. They have described Conectado and presents its validation with actual students using game analytics.
- Krishanu Maity, et al., [9] investigated the role of sentiment and emotion information in identifying cyberbullying in the Indian scenario. From Twitter, a benchmark Hind–English code-mixed corpus called BullySentEmo has been developed as there is no dataset available labeled with bully, sentiment, and emotion. An attention-based multimodal, adversarial multitasking framework is proposed for cyberbully detection (CBD) considering two auxiliary tasks: sentiment analysis (SA) and emotion recognition(ER).

Limitations of Existing System

The existing approaches focus on how offensive the content of the message is based on that they identify cyberbullies but does not consider why the message was offensive, i.e., they do not analyse the context of the entire conversation just the content of the message. Our approach utilizes the bag-of-words with the text to identify curse words and use SA to determine the emotions or attitude of the sender, and finally, we analyse the entire context in which the sender and receiver communicate. These overlooked factors could significantly or completely change the results of cyberbullying detection.

Proposed System

The proposed system contains a three-phase algorithm, called BullyNet, for detecting cyberbullies on Twitter social network which exploits bullying tendencies by proposing a robust method for constructing a cyberbullying signed network (SN). And analyze tweets to determine their relation to cyberbullying while considering the context in which the tweets exist in order to optimize their bullying score. Also proposed a centrality measure to detect cyberbullies from a cyberbullying SN and show that it outperforms other existing measures.

Architecture

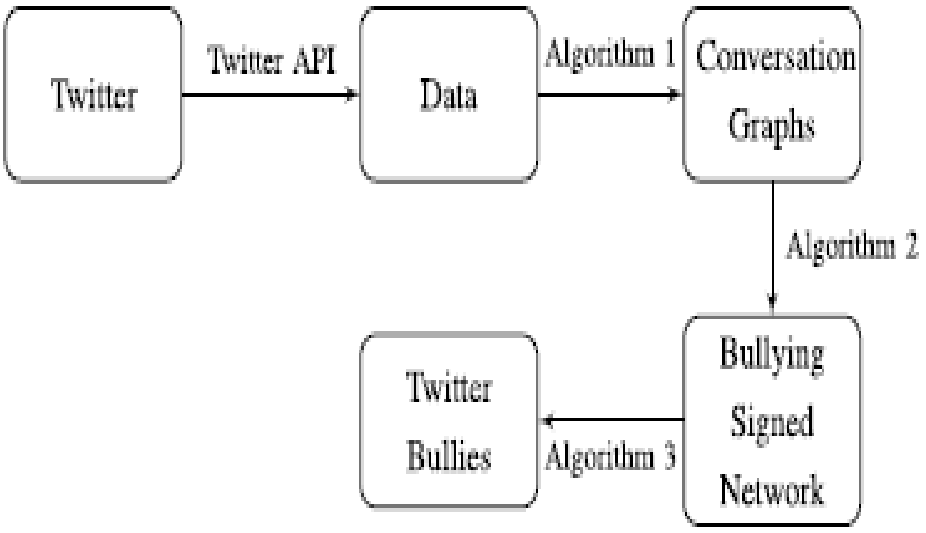

Fig. 1 Protocol flowchart of BullyNet

In this section, we first present an overview of the proposed three-phase bully finding algorithm (BFA) and elaborate the steps in each phase. The objective of our solution is to identify the bullies from raw Twitter data based on the context as well as the contents in which the tweets exist. Given a set of tweets *T* containing Twitter features such as user ID, reply ID, and so on, the proposed approach consists of three algorithms: 1) conversation graph generation algorithm; 2) bullying SN generation algorithm; and 3) BFA. The first algorithm constructs a directed weighted conversation graph *Gc* by efficiently reconstructing the conversations from raw Twitter data while enabling a more accurate model of human interactions. The second algorithm constructs a bullying SN *B* to analyze the behavior of users in social media. The third algorithm consists of our proposed A&M centrality measures to identify bullies from *B*. Fig. 1 shows the process flow of BullyNet where the raw data are extracted from Twitter using Twitter API from which the conversation graph is constructed for each conversation using Algorithm 1. Then, from the conversation graphs, a bullying SN is generated using Algorithm 2. Finally, the bullies from Twitter are identified by applying Algorithm 3.

Methodologies/Algorithms

- The first algorithm constructs a directed weighted conversation graph Gc by efficiently reconstructing the conversations from raw Twitter data while enabling a more accurate model of human interactions.
- The second algorithm constructs a bullying SN B to analyze the behavior of users in social media.
- The third algorithm consists of our proposed A&M centrality measures to identify bullies from B.

Algorithm 1 Conversation Graph Generation

Input: Set of tweets, $T = \{t_1, \ldots, t_n\}$
Output: Conversation graphs $G_c = \{g_{c_1}, \ldots, g_{c_m}\}$

1) Sort all tweets in T in reverse-chronological order based on date of creation.
2) For each tweet t_i in T, where $1 \leq i \leq |T|$:
   a) If t_i does not belong to a conversation, then create a new conversation $c \in C$ and associate t_i with c.
   b) If there is a tweet $t' \in \{t_i, t_{i+1}, \ldots, t_{|T|}\}$ where $\mathcal{DID}(t_i) = \mathcal{SID}(t')$ then associate t' with all t_i's conversations.
3) For each conversation $c_i \in C$:
   a) Construct a conversation graph $g_{c_i} \in G_c$, where users are represented as nodes and tweets as edges.
   b) For each edge $e = (u, v)$ in g_{c_i}:
      i) Compute the sentiment of the tweet (SA).
      ii) Compute the cosine similarity (CS) of the tweet with bullying bag of words (CS).
      iii) Calculate the bullying indicator I_{t_i} (weight) of the edge as follows:
      $I_{uv} = \beta * SA + \gamma * CS$
4) Return G_c

Algorithm 2 Bullying SN Generation

Input: Set of conversation graphs, G_c
Output: Bullying Signed Network $\mathcal{B}$

1) For each conversation graph g_{c_i} in G_c:
   a) For each set of edges with the same order, sorted ascendingly, compute the bullying score of source node u toward target node v for each edge $e = (u, v)$ as follows:
   $S_{uv} = I_{uv} + \alpha(I_{uv} - S_{vu})$.
   and then determine the average score of node u for the same set of edges.
   b) Compute the overall bullying score S of each node in g_{c_i} as follows:
      i) If the node is the *root* node, then: $S = \frac{\sum S}{1+2.2(n-1)}$
      ii) Otherwise: $S = \frac{\sum S}{2.2(n)}$
2) Construct the bullying SN graph $\mathcal{B}$ by merging all the conversation graphs together.
3) Return $\mathcal{B}$.

Algorithm 3 BFA

Input: Bullying Signed Network $G_s = (V, E, W)$
Output: List of bullies and its attitude score $L = [(u_1, s_1), (u_2, s_2), \ldots, (u_{|L|}, s_{|L|})]$

1) Initialize $M^0(v) = -1$ and $A^0(v) = -1, \forall v \in V$.
2) Set iteration index i = 1
   a) For each $v \in V$ compute merit score
   $M^i(v) = \frac{1}{2|in(v)|} \sum_{u \in in(v)} (w_{uv})(A^{i-1}(u))$ where $|in(v)|$ is the number of incoming edges to the node v
   b) For each $u \in V$ compute attitude score
   $A^i(u) = \frac{1}{2|out(u)|} \sum_{v \in out(u)} (w_{uv} + X_{uv})$ where $|out(u)|$ is the number of outgoing edges from the node u
3) If there exist atleast one $v \in V : M^i(v) \neq M^{i-1}(v)$ or $A^i(v) \neq A^{i-1}(v)$
   a) Increase the iteration index $i = i + 1$
   b) Repeat step 2a & 2b for each iteration
4) For each $v \in V$ add the node and its corresponding attitude score value greater than 0 to the list L
5) Return L

Applications

- By using BullyNet, tweets with malicious intent can be detected and take actions on the people who are using social media like twitter to spread hate amongst the people.
- It provide the security to people.

References

[1] Wan Noor Hamiza Wan Ali, Masnizah Mohd, Fariza Fauzi, “Cyberbullying Detection: An Overview,” 2018 Cyber Resilience Conference (CRC).

[2] Djedjiga Mouheb, Masa Hilal Abushamleh, Maya Hilal Abushamleh, Zaher Al Aghbari, Ibrahim Kamel, “ Real-time Detection of Cyberbullying in Arabic Twitter Streams, ” 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS).

[3] Vijay Banerjee, Jui Telavane, Pooja Gaikwad, Pallavi Vartak, “Detection of Cyberbullying Using Deep Neural Network, ” 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).

[4] Vivek K. Singh, Connor Hofenbitzer, “Fairness across Network Positions in Cyberbullying Detection Algorithms, ” 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[5] Ong Chee Hang, Halina Mohamed Dahlan, “Cyberbullying Lexicon for Social Media, ” 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS).

[6] Jason Wang, Kaiqun Fu, Chang-Tien Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection, ” 2020 IEEE International Conference on Big Data (Big Data).

[7] Ankit Pradhan, Venu Madhav Yatam, Padmalochan Bera, “Self-Attention for Cyberbullying Detection, ” 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA).

[8] Antonio Calvo-Morata, Dan Cristian Rotaru, Cristina Alonso-Fernández, Manuel Freire-Morán, Iván Martínez-Ortiz, Baltasar Fernández-Manjón, “Validation of a Cyberbullying Serious Game Using Game Analytics , ” 2020 IEEE Transactions on Learning Technologies.

[9] Krishanu Maity, Sriparna Saha, Pushpak Bhattacharyya, “Emoji, Sentiment and Emotion Aided Cyberbullying Detection in Hinglish, ” 2022 IEEE Transactions on Computational Social Systems.