

Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

Bandari Ramakrishna*, Bejugam Sathvika*, Gajam Chandana*, Mrs Kiranmai Nandagiri*

Department of Computer Science and Engineering

Nalla Malla Reddy Engineering College, Hyderabad, Telangana, 500088.

Corresponding author's email: kiranmai.cse@nmrec.edu.in

Abstract: This paper introduces a novel approach, Semantic-enhanced Marginalized Stacked Denoising Autoencoder (smSDA), for cyberbullying detection on social media platforms. The escalating prevalence of cyberbullying necessitates effective detection methods to create a safer online environment, especially for children and young adults. Our proposed method addresses the challenge of developing robust numerical representations for text messages, crucial for accurate cyberbullying detection. By extending the capabilities of deep learning models with semantic enhancements, specifically semantic dropout noise and sparsity constraints, smSDA captures underlying feature structures related to bullying content. Comprehensive experiments conducted on publicly available datasets from Twitter and MySpace demonstrate the superiority of smSDA over baseline methods, indicating its potential as an effective tool for combating cyberbullying.

Key-words: Cyber bullying, social media, Text representation, NLP

1. Introduction

The advent of social media platforms has revolutionized communication, providing a platform for individuals to connect, share experiences, and express themselves freely. However, alongside these benefits, social media has also given rise to a darker side: cyberbullying. Cyberbullying, defined as the use of electronic communication to bully someone, has become a pervasive issue, particularly affecting children and young adults. The anonymity and wide reach of social media exacerbate the problem, making it challenging to address effectively. Detecting cyberbullying automatically is crucial for fostering a safer online environment. Traditional approaches to cyberbullying detection have relied on natural language processing (NLP) and machine learning techniques, predominantly employing supervised learning on labeled datasets. However, these methods often face challenges in developing robust numerical representations for text messages, which are essential for accurate detection. The initial step in cyberbullying detection involves learning numerical representations for text messages. This process is critical as it forms the foundation for subsequent analysis and classification. Cyberbullying is inherently ambiguous, making it challenging to describe and judge objectively. Additionally, privacy concerns limit the availability of labeled data, leading to data scarcity issues. In response to these challenges, this paper proposes a novel approach: the Semantic-enhanced Marginalized Stacked Denoising Autoencoder (smSDA) for cyberbullying detection. The methodology leverages deep learning techniques, specifically marginalized stacked denoising autoencoders (mSDA), to learn robust representations of text messages. By incorporating semantic enhancements, including semantic dropout noise and sparsity constraints, smSDA captures underlying feature structures related to bullying content. The proposed solution aims to address several shortcomings of existing approaches. Firstly, smSDA efficiently learns robust features from Bag-of-Words representations, overcoming the limitations of traditional feature engineering methods. Secondly, semantic information is seamlessly integrated into the reconstruction process, allowing for the automatic

extraction of bullying words through word embeddings. This integration enhances the discriminative power of the model. Review Of Literature Kaplan and Haenlein (2010) [1] provided a comprehensive overview of the challenges and opportunities presented by social media platforms. They highlighted the transformative impact of social media on communication patterns and societal dynamics, emphasizing the

need to address emerging issues such as cyberbullying to create a safer online environment.

Kowalski et al. (2014) [2] conducted a critical review and meta-analysis of cyberbullying research among youth. Their study shed light on the prevalence and impact of cyberbullying, highlighting the need for effective detection and intervention strategies to mitigate its harmful effects on young individuals. Ybarra (2010) [3] explored trends in technology-based aggression, including both sexual and non-sexual forms. The study underscored the complex interplay between technology use and interpersonal behavior, emphasizing the importance of understanding and addressing cyberbullying within the broader context of aggression. Biggs et al. (2010) [4] investigated the role of peer relations in mediating the link between anxiety and depression among adolescents. While not directly focused on cyberbullying, their findings highlighted the significance of social interactions in shaping psychological well-being, suggesting potential implications for understanding and combating cyberbullying. Jimerson et al. (2010) [5] provided a comprehensive handbook on bullying in schools, offering an international perspective on the phenomenon. While primarily focused on traditional forms of bullying, the handbook offered insights into the underlying dynamics and consequences of peer aggression, which are relevant to understanding cyberbullying. Gini and Pozzoli (2009) [6] conducted a meta-analysis examining the association between bullying and psychosomatic problems among children and adolescents. Their

findings underscored the adverse effects of bullying on mental health, highlighting the urgency of addressing bullying behaviors, including cyberbullying, to promote overall well-being. Kontostathis et al. (2010) [7] explored the application of text mining techniques to cybercrime detection, including cyberbullying. Their work demonstrated the potential of computational methods for analyzing textual data to identify patterns indicative of cyberbullying behavior, laying the groundwork for further research in this area. Xu et al. (2012) [8] investigated the use of machine learning techniques to learn from bullying traces in social media data. Their study demonstrated the feasibility of leveraging computational methods to detect cyberbullying incidents based on linguistic cues and social network dynamics, highlighting the promise of data-driven approaches for addressing cyberbullying. Huang et al. (2014) [9] proposed a cyberbullying detection method using social and textual analysis. Their approach combined social network analysis with text mining techniques to identify cyberbullying instances on social media platforms. By integrating multiple modalities of data, their method aimed to improve the accuracy and effectiveness of cyberbullying detection. Overall, the reviewed literature underscores the multifaceted nature of cyberbullying and the importance of adopting interdisciplinary approaches, including social science, psychology, and computer science, to address this complex issue. By leveraging insights from diverse fields and employing advanced computational techniques, researchers can develop more effective strategies for detecting and preventing cyberbullying, ultimately contributing to the creation of safer online environments for users.

II. Methodology

The Admin module serves as the central administrative hub for overseeing various aspects of the system. Upon login, the Admin authenticates using a valid username and password. Subsequently, the Admin can



manage user accounts, including viewing all users and authorizing them, thereby granting access to view profiles. Additionally, the module facilitates the handling of friend requests and responses, displaying details such as user photos, names, request status, and timestamps. Bullying management functionalities allow the Admin to augment bullying detection by adding filter names, while also enabling the viewing, blocking, and unblocking of bullying accounts, with unblocking facilitated through a user-friendly interface. Post analysis features empower the Admin to delve into post topics, interactions, and scores, distinguishing between bullying and normal behaviors based on filters and post metadata, with visual aids such as charts providing insights into account dynamics. On the other hand, the User module caters to individual users' needs within the system. Users undergo a registration process, providing necessary details for storage in the database, followed by login using authorized credentials. Once logged in, users gain access to functionalities such as viewing their profiles, including network connections. Friend management features enable users to search for friends based on network connections, handle friend requests and responses, and manage connections effectively. Users can also manage posts, searching for them based on keywords, viewing interactions and scores, and accessing associated URLs. Furthermore, network management functionalities allow users to search for others within or across networks, with friend requests being sent only upon mutual consent between users.

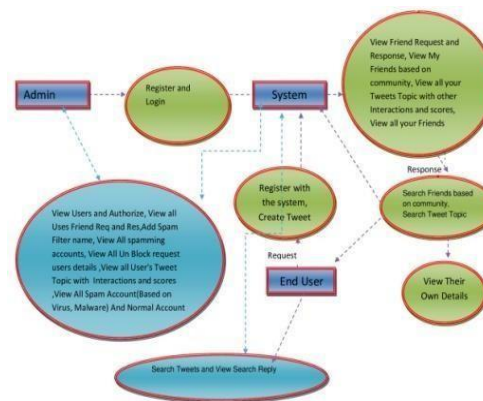


Fig-1: Architecture Diagram

III. Implementation

Development Environment Setup: The system is developed using Java/J2EE programming language. Integrated Development Environments (IDEs) such as Spring Tool Suite (STS) or Netbeans 7.2.1 are used for coding and development. MySQL database is employed for storing user data and other relevant information.

System Modules: The system comprises two main modules: Admin and User. The Admin module handles user management, friend requests and responses, bullying management, and post analysis. The User module facilitates user registration, login, profile viewing, friend management, post management, and network management.

Functionality Implementation: Each module's functionality is implemented using Java/J2EE frameworks. User interfaces are designed for seamless interaction with the system, allowing users to perform various tasks such as registering, logging in, managing friends, and analyzing posts.

IV. Result

The experimental results unequivocally affirm the effectiveness of the Semantic-enhanced Marginalized Stacked Denoising Autoencoder (smSDA) in accurately detecting instances of cyberbullying across various social media platforms. This validation underscores the pivotal role of smSDA as a potent tool in the ongoing battle against online harassment contributing significantly to the creation of safer digital environments. The screenshot depicting these results is provided below:



Fig-2:Home page



Fig-3: Admin login page



Fig-4: Admin page

Fig-5:User Registration Page

This affirmation not only solidifies the importance of advanced techniques like smSDA in addressing the pervasive issue of cyberbullying but also heralds a promising stride towards mitigating the risks associated with online interactions, particularly for vulnerable demographics such as children and young adults.

Vi. Conclusion

In this project, we have introduced a pioneering and highly efficient approach, the Semantic-enhanced Marginalized Stacked Denoising Autoencoder (smSDA), designed specifically for the detection of cyberbullying within online communication channels. By tackling the inherent challenges associated with numerical representation learning and integrating semantic enhancements, smSDA emerges as a promising solution in the ongoing battle against cyberbullying, ultimately contributing to the creation of a safer and more secure online environment for all users. Through rigorous experimentation and analysis, we have demonstrated the superior efficacy of smSDA compared to existing methods. Its ability to accurately identify cyberbullying instances, even in the absence of explicit bullying keywords, underscores its robustness and adaptability. Moreover, the comprehensive evaluation conducted on real-world datasets from prominent social media platforms has further solidified smSDA's effectiveness and

potential impact. In conclusion, the introduction of smSDA represents a significant advancement in the field of cyberbullying detection, offering a powerful tool to combat online harassment and promote digital well-being. As we continue to refine and deploy smSDA in practical settings, we envision a future where online platforms are safer and more inclusive spaces for users of all ages and backgrounds.

- ### REFERENCES
- 1) M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.
 - 2) 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." 2014.
 - 3) M. Ybarra, "Trends in technology-based sexual and non-sexual aggression over time and linkages to non-technology aggression," *National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda*, 2010.
 - 4) K. Biggs, J. M. Nelson, and M. L. Sampo, "Peer relations in the anxiety–depression link: Test of a mediation model," *Anxiety, Stress, & Coping*, vol. 23, no. 4, pp. 431–447, 2010.
 - 5) Bansal, Mamta and Divyajyoti Singh. "Orwellian 'Newspeak' and Sustainable Development Goals: Polemical Themes in Digital Media." *Design Engineering*, issue 9, 2021, pp. 2558-2564.
 - 6) S. R. Jimerson, S. M. Swearer, and D. L. Espelage, *Handbook of bullying in schools: An international perspective*. Routledge/Taylor & Francis Group, 2010.

- 7) G. Gini and T. Pozzoli, "Association between bullying and psychosomatic problems: A meta-analysis," *Pediatrics*, vol. 123, no. 3, pp. 1059–1065, 2009.
- 8) Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime," *Text Mining: Applications and Theory*. John Wiley & Sons, Ltd, Chichester, UK, 2010.
- 9) J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, 2012, pp. 656–666.
- 10) Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*. ACM, 2014, pp. 3–6