A survey on Semantic-Enhanced Marginalized Denoising Auto-Encoder

S MANIKANTA[1],ABDUL AHAD[2], DDD SURIBABU[3]
[1]PG Scholar In Department of CSE DNR College of Engineering &Technology ,Bhimavaram ,A.P
[2] Assoc. Prof. In Department of CSE DNR College of Engineering &Technology ,Bhimavaram ,A.P
[3]HEAD & Assoc.Prof In Department of CSE DNR College of Engineering &Technology ,Bhimavaram ,A.P

Abstract—As a side effect of increasingly popular social media, cyberbullying has emerged as a serious problem afflicting children, adolescents and young adults. Machine learning techniques make automatic detection of bullying messages in social media possible, and this could help to construct a healthy and safe social media environment. In this meaningful research area, one critical issue is robust and discriminative numerical representation learning of text messages. In this paper, we propose a new representation learning method to tackle this problem. Our method named Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) is developed via semantic extension of the popular deep learning model stacked denoising autoencoder. The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. Our proposed method is able to exploit the hidden feature structure of bullying information and learn a robust and discriminative representation of text. Comprehensive experiments on two public cyberbullying corpora (Twitter and MySpace) are conducted, and the results show that our proposed approaches outperform other baseline text representation learning methods.

Keywords—Cyberbullying Detection, Text Mining, Representation Learning, Stacked Denoising Autoencoders

I. INTRODUCTION

Cyberbullying can be defined as aggressive, intentional actions performed by an individual or a group of people via digital communication methods such as sending messages and posting comments against a victim. Different from traditional bullying that usually occurs at school during face-to-face communication, cyberbullying , cyberbullying victimization rate ranges from 10% to 40%. In the United States, approximately 43% of teenagers were ever bullied on social media . The same as traditional bullying, cyberbullying has negative, insidious and sweeping impacts on children. The outcomes for victims under cyberbullying may even be tragic such as the occurrence of self-injurious behaviour or suicides. One way to address the cyberbullying problem isto automatically detect and promptly report bullying messages so that proper measures can be taken to prevent possi-ble tragedies.

In the text-based cyberbullying detection, the first and also critical step is the numerical representation learning for text messages. In fact, representation learning of text is extensively studied in text mining, information retrieval and natural language processing (NLP). Bag-of-words (BoW) model is one commonly used model that each dimension corresponds to a term. Latent Semantic Analysis (LSA) and topic models are another popular text representation models, Therefore, the useful representation should discover the meaning behind text units. In cyberbullying detection, the numerical representation for Internet mes-sages should be robust and discriminative. Since messages on social media are often very short and contain a lot of informal language and misspellings, robust representations for these messages are required to reduce their ambigui-ty. Even worse, the lack of sufficient high-quality training data, i.e., data sparsity make the issue more challenging. Firstly, labeling data is labor intensive and time consuming. Secondly, cyberbullying is hard to describe and judge from a third view due to its intrinsic ambiguities. Thirdly, due to protection of Internet users and privacy issues, only a small portion of messages are left on the Internet, and most bullying posts are deleted. As a result, the trained classifier may not generalize well on testing messages that contain nonactivated but discriminative features. The goal of this present study is to develop methods that can learn robust and discriminative representations to tackle the above prob-lems in cyberbullying detection Some approaches have been proposed to tackle these problems by incorporating expert knowledge into feature learning. Yin et.al proposed to combine BoW features, senti-ment features and contextual features to train a support vec-tor machine for online harassment detection . Dinakar et.al utilized label specific features to extend the general features, where the label specific features are learned by Linear Discriminative Analysis . In addition, common sense knowledge was also applied. Nahar et.al presented a weighted TF-IDF scheme via scaling bullying-like features by a factor of two . Besides content-based information, Maral et.al proposed to apply users' information, such as gender and history messages, and context information as extra features .

Figure:Cyberbulling victimization rates

But a major limitation of these approaches is that the learned feature space still relies on the BoW assumption and may not be robust. In addition, the performance of these approaches rely on the quality of hand-crafted features, which require extensive domain knowledge. there is a strong correlation between bullying word fuck and normal word off since they often occur together. If bullying messages do not contain such obvious bullying features, such as fuck is often misspelled as fck, the correlation may help to reconstruct the bullying features from normal ones

addition, L1 regularization of the projection matrix is added to the objective function of each autoencoder layer in our model to enforce the sparstiy of projection matrix, and this in turn facilitates the discovery of the most relevant terms for reconstructing bullying terms. The main contributions of our work can be summarized as follows:

* Semantic information is incorporated into the re-construction and modifications.Finally, these specialized modifications make the new feature space more discriminative and this in turn facilitates bullying detection.
* Comprehensive experiments on real-data sets have verified the performance of our proposed model.

II. oBJECTIVE

In text mining, information retrieval and natural language processing, effective numerical representation of linguistic units is a key issue. cyberbullying has emerged as a serious problem afflicting children and young adults. Previous studies of cyberbullying focused on extensive surveys and its psy-chological effects on victims, and were mainly conducted by social scientists and psychologists.

Figure:Cyberbullying victimization

cy-berbullying, the psychological science approach based on personal surveys is very time-consuming and may not be suitable for automatic detection of cyberbullying. Since machine learning is gaining increased popularity in recent years, the computational study of cyberbullying has at-tracted the interest of researchers. Several research areas including topic detection and affective analysis are closely related to cyberbullying detection. Owing to their efforts, automatic cyberbullying detection is becoming possible. In machine learning-based cyberbullying detection, there are two issues: 1) text representation learning to transform each post/message into a numerical vector and 2) classifier train-ing. Xu et.al presented several off-the-shelf NLP solutions including BoW models, LSA and LDA for representation learning to capture bullying signals in social media [8]. As an introductory work, they did not develop specialized models for cyberbullying detection. The performance of label-specific features largely depends on the size of training corpus. In addition, they need to construct a bullyspace knowledge base to boost the performance of natural language processing methods

III. PROPOSAL

A bullying trace is defined as the response of participants to their bullying experience. Bullying traces include not only mes-sages about direct bullying attack, but also messages about reporting a bullying experience, revealing self as a victim et. al. Therefore, bullying traces far exceed the incidents of cyberbullying. Automatic detection of bullying traces are valuable for cyberbullying The MySpace dataset is crawled from MySpace groups. Each group consists of several posts by different users, which can be regarded as a conversation about one topic. Due to the interactive nature behind cyberbullying, The raw text for these data, as XML files, have been kindly provided by Kontostathis et.al[3]. The XML files contain information about the posts, such as post text, post data, and users' information. Since there were no standard splits of training vs. test datasets in our adopted Twitter and MySpace corpora, we need to define the training and testing datasets. As analyzed above that the lack of labeled training corpus hinders the development of automatic cyberbullying detection, the sizes of training corpus are all controlled to be very small in our experiments. To reduce variance, the process is repeated ten times so that we can have ten sub-datasets from Twitter data. For MySpace dataset, we also randomly pick 400 data samples as the training corpus and use the rest data for all features used. It is difficult to learn robust features for small training data by intensifying each bullying features' amplitude. Our approach aims to find the cor-relation between normal features and bullying features by reconstructing corrupted data so as to yield robust features. Comparing the performances of smSDA and $smSDA_u$, which adopt biased semantic dropout noise and unbi-ased semantic dropout noise, respectively. The results have shown that $smSDA_u$ performs slightly worse than smSDA. This may be explained by the fact that the unbiased semantic dropout noise cancels the enhancement of bullying features.

.

IV. CONCLUSIONS

This paper addresses the text-based cyberbullying detection problem, where robust and discriminative representations of messages are critical for an effective detection system. By designing semantic dropout noise and enforcing sparsity, we have developed semantic-enhanced marginalized denoising autoencoder as a specialized representation learning process.Experimentally verified through two cyberbullying frames.

REFERENCES

[1] A. 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 nontechnology aggression," National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda, 2010.

[4] B. K. Biggs, J. M. Nelson, and M. L. Sampilo, "Peer relations in the anxiety–depression link: Test of a mediation model," Anxiety, Stress, & Coping, vol. 23, no. 4, pp. 431–447, 2010.

[5] S. R. Jimerson, S. M. Swearer, and D. L. Espelage, Handbook of bullying in schools: An international perspective. Routledge/Taylor & Francis Group, 2010.

[6] G. Gini and T. Pozzoli, "Association between bullying and psy-chosomatic problems: A meta-analysis," Pediatrics, vol. 123, no. 3, pp. 1059–1065, 2009.

[7] A. Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime," Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, 2010.

[8] 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.

[9] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in Proceedings of the 3rd Inter-national Workshop on Socially-Aware Multimedia. ACM, 2014, pp. 3–6.

[10] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7, 2009.

[11] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detec-tion of textual cyberbullying." in The Social Mobile Web, 2011.

[12] V. Nahar, X. Li, and C. Pang, "An effective approach for cy-berbullying detection," Communications in Information Science and Management Engineering, 2012.

[13] M. Dadvar, F. de Jong, R. Ordelman, and R. Trieschnigg, "Im-proved cyberbullying detection using gender information," in Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop (DIR2012). Ghent, Belgium: ACM, 2012.

[14] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Im-proving cyberbullying detection with user context," in Advances in Information Retrieval. Springer, 2013, pp. 693–696.

[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," The Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.

[16] P. Baldi, "Autoencoders, unsupervised learning, and deep archi-tectures," Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7, p. 43, 2012.

[17] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized de-noising autoencoders for domain adaptation," arXiv preprint arX-iv:1206.4683, 2012.

[18] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse processes, vol. 25, no. 2-3, pp. 259–284, 1998.

[19] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceed-ings of the National academy of Sciences of the United States of America, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.

[21] T. Hofmann, "Unsupervised learning by probabilistic latent se-mantic analysis," Machine learning, vol. 42, no. 1-2, pp. 177–196, 2001.

[22] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 8, pp. 1798–1828, 2013.

[23] B. L. McLaughlin, A. A. Braga, C. V. Petrie, M. H. Moore et al., Deadly Lessons:: Understanding Lethal School Violence.

National Academies Press, 2002.

[24] J. Juvonen and E. F. Gross, "Extending the school ground-s?bullying experiences in cyberspace," Journal of School health, vol. 78, no. 9, pp. 496–505, 2008.

[25] M. Fekkes, F. I. Pijpers, A. M. Fredriks, T. Vogels, and S. P. Verloove-Vanhorick, "Do bullied children get ill, or do ill chil-dren get bullied? a prospective cohort study on the relationship between bullying and health-related symptoms," Pediatrics, vol. 117, no. 5, pp. 1568–1574, 2006.

[26] M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, "Brute force works best against bullying," in Proceedings of IJCAI 2015 Joint Workshop on Constraints and Preferences for Configuration and Recommendation and Intelligent Techniques for Web Personaliza-tion. ACM, 2015.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society. Series B (Methodological), pp. 267–288, 1996.

[28] C. C. Paige and M. A. Saunders, "Lsqr: An algorithm for sparse linear equations and sparse least squares," ACM Transactions on Mathematical Software (TOMS), vol. 8, no. 1, pp. 43–71, 1982.

[29] M. A. Saunders et al., "Cholesky-based methods for sparse least squares: The benefits of regularization," Linear and Nonlinear Con-jugate Gradient-Related Methods, pp. 92–100, 1996.

[30] J. Fan and R. Li, "Variable selection via nonconcave penalized like-lihood and its oracle properties," Journal of the American statistical Association, vol. 96, no. 456, pp. 1348–1360, 2001.