International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

G Netaji

B V C Engineering College, Odalarevu Andhra Pradesh, India

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 : Detection, Cyberbullying,, Social Networking, Denoising

I. INTRODUCTION

Social Media, as defined in [1], is ‘’a group of Internetbased applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.‘’ Via social media, people can enjoy enormous information, convenient communication experience and so on. However, social media may have some side effects such as cyberbullying, which may have negative impacts on the life of people, especially children and teenagers. 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 facetoface communication, cyberbullying on social media can take place anywhere at any time. For bullies, they are free to hurt their peers’ feelings because they do not need to face someone and can hide behind the Internet. For victims, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or social media. As reported in [2], cyberbullying victimization rate ranges from 10% to 40%. In the United States, approximately 43% of teenagers were ever bullied on social media [3]. The same as traditional bullying, cyberbullying has negative, insidious and sweeping impacts on children [4], [5], [6]. The outcomes for victims under cyberbullying may even be tragic such as the occurrence of self-injurious behavior or suicides.

One way to address the cyberbullying problem is to automatically detect and promptly report bullying messages so that proper measures can be taken to prevent possible tragedies. Previous works on computational studies of bullying have shown that natural language processing and machine learning are powerful tools to study bullying [7], [8]. Cyberbullying detection can be formulated as a supervised learning problem. A classifier is first trained on a cyberbullying corpus labeled by humans, and the learned classifier is then used to recognize a bullying message. Three kinds of information including text, user demography, and

social network features are often used in cyberbullying detection [9]. Since the text content is the most reliable, our work here focuses on text-based cyberbullying detection.

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, which are both based on BoW models. By mapping text units into fixed-length vectors, the learned representation can be further processed for numerous language processing tasks. Therefore, the useful representation should discover the meaning behind text units. In cyberbullying detection, the numerical representation for Internet messages 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 ambiguity.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 problems in cyberbullying detection.

In this paper, we investigate one deep learning method named stacked denoising autoencoder (SDA) [15]. SDA stacks several denoising autoencoders and concatenates the output of each layer as the learned representation. Each denoising autoencoder in SDA is trained to recover the input data from a corrupted version of it. The input is corrupted by randomly setting some of the input to zero, which is called dropout noise. This denoising process helps the autoencoders to learn robust representation. In addition, each autoencoder layer is intended to learn an increasingly abstract representation of the input.
An automatic extraction of bullying words based on word embeddings is proposed so that the involved human labor can be reduced. During training of smSDA, we attempt to reconstruct bullying features from other normal words by discovering the latent structure, i.e. correlation, between bullying and normal words. The intuition behind this idea is that some bullying messages do not contain bullying words. The correlation information discovered by smSDA helps to reconstruct bullying features from normal words, and this in turn facilitates detection of bullying messages without containing bullying words.

II. METHODOLOGY

Previous works on computational studies of bullying have shown that natural language processing and machine learning are powerful tools to study bullying. Cyberbullying detection can be formulated as a supervised learning problem. A classifier is first trained on a cyberbullying corpus labeled by humans, and the learned classifier is then used to recognize a bullying message. Yin et.al proposed to combine BoW features, sentiment features and contextual features to train a support vector 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. The first and also critical step is the numerical representation learning for text messages. 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.

Three kinds of information including text, user demography, and social network features are often used in cyberbullying detection. Since the text content is the most reliable, our work here focuses on text-based cyberbullying detection.

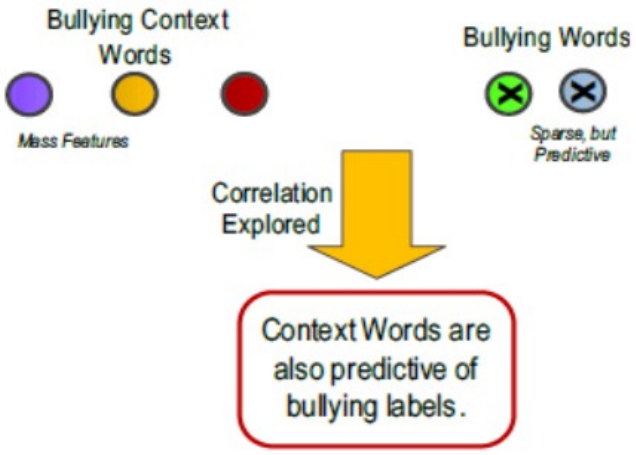
In this paper, we investigate one deep learning method named stacked denoising autoencoder (SDA). SDA stacks several denoising autoencoders and concatenates the output of each layer as the learned representation. Each denoising autoencoder in SDA is trained to recover the input data from a corrupted version of it. The input is corrupted by randomly setting some of the input to zero, which is called dropout noise. This denoising process helps the autoencoders to learn robust representation. In addition, each autoencoder layer is intended to learn an increasingly abstract representation of the input.

In this paper, we develop a new text representation model based on a variant of SDA: marginalized stacked denoising autoencoders (mSDA), which adopts linear instead of nonlinear projection to accelerate training and marginalizes infinite noise distribution in order to learn more robust representations. We utilize semantic information to expand mSDA and develop Semantic-enhanced Marginalized Stacked Denoising Autoencoders (smSDA). The semantic information consists of bullying words. An automatic extraction of bullying words based on word embeddings is proposed so that the involved human labor can be reduced. During training of smSDA, we attempt to reconstruct bullying features from other normal words by discovering the latent structure, i.e. correlation, between bullying and normal words. The intuition behind this idea is that some bullying messages do not contain bullying words. The correlation information discovered by smSDA helps to reconstruct bullying features from normal words, and this in turn facilitates detection of bullying messages without containing bullying words.

Our proposed Semantic-enhanced Marginalized Stacked Denoising Autoencoder is able to learn robust features from BoW representation in an efficient and effective way. These robust features are learned by reconstructing original input from corrupted (i.e., missing) ones. The new feature space can improve the performance of cyberbullying detection even with a small labeled training corpus. Semantic information is incorporated into the reconstruction process via the designing of semantic dropout noises and imposing sparsity constraints on mapping matrix. In our framework, high-quality semantic information, i.e., bullying words, can be extracted automatically through word embeddings.

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.

1. Semantic-Enhanced Marginalized Denoising Auto-Encoder



The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

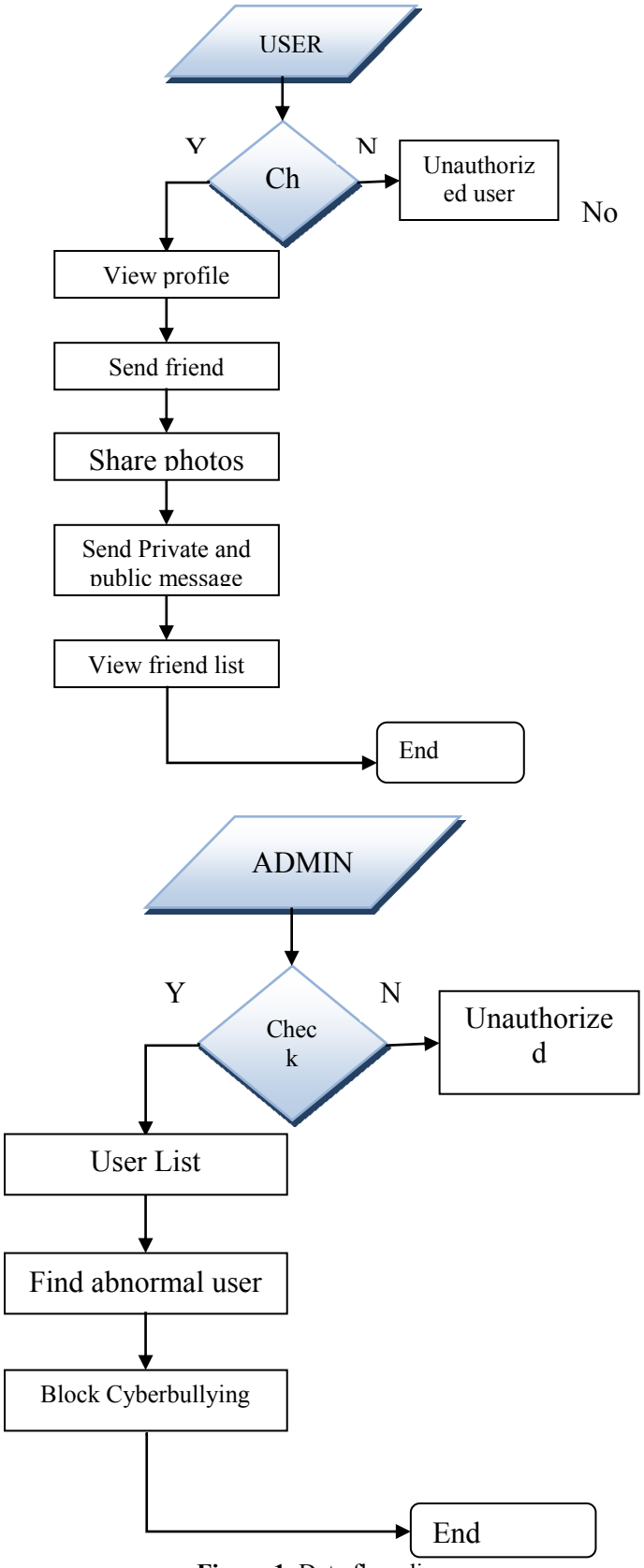


Figure 1. Data flow diagram

III. IMPLEMENTATION

A. OSN System Construction Module

In the first module, we develop the Online Social Networking (OSN) system module. We build up the system with the feature of Online Social Networking. Where, this module is used for new user registrations and after registrations the users can login with their authentication. Where after the existing users can send messages to privately and publicly, options are built. Users can also share post with others. The user can able to search the other user profiles and public posts. In this module users can also accept and send friend requests. With all the basic feature of Online Social Networking System modules is build up in the initial module, to prove and evaluate our system features.

B. Construction of Bullying Feature Set:

The bullying features play an important role and should be chosen properly. In the following, the steps for constructing bullying feature set Zb are given, in which the first layer and the other layers are addressed separately. For the first layer, expert knowledge and word embeddings are used. For the other layers, discriminative feature selection is conducted.

In this module firstly, we build a list of words with negative affective, including swear words and dirty words. Then, we compare the word list with the BoW features of our own corpus, and regard the intersections as bullying features. Finally, the constructed bullying features are used to train the first layer in our proposed smSDA. It includes two parts: one is the original insulting seeds based on domain knowledge and the other is the extended bullying words via word embeddings. Observe Attentively Over A Period Of Time.

C. Cyberbullying Detection

In this module we propose the Semantic-enhanced Marginalized Stacked Denoising Auto-encoder (smSDA). In this module, we describe how to leverage it for cyberbullying detection. smSDA provides robust and discriminative representations The learned numerical representations can then be fed into our system. In the new space, due to the captured feature correlation and semantic information, even trained in a small size of training corpus, is able to achieve a good performance on testing documents.Based on word embeddings, bullying features can be extracted automatically. In addition, the possible limitation of expert knowledge can be alleviated by the use of word embedding.

D. Semantic-Enhanced Marginalized Denoising Auto-Encoder:

An automatic extraction of bullying words based on word embeddings is proposed so that the involved human labor can be reduced. During training of smSDA, we attempt to reconstruct bullying features from other normal words by discovering the latent structure, i.e. correlation, between bullying and normal words. The intuition behind this idea is that some bullying messages do not contain bullying words.

The correlation information discovered by smSDA helps to reconstruct bullying features from normal words, and this in turn facilitates detection of bullying messages without containing bullying words. For example, 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 so that the bullying message can be detected. It should be noted that introducing dropout noise has the effects of enlarging the size of the dataset, including training data size, which helps alleviate the data sparsity problem.

IV. CONCLUSION

In, 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 model for cyberbullying detection. In addition, word embeddings have been used to automatically expand and refine bullying word lists that is initialized by domain knowledge. The performance of our approaches has been experimentally verified through two cyberbullying corpora from social medias: Twitter and MySpace. As a next step we are planning to further improve the robustness of the learned representation by considering word order in messages.

V. 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 metaanalysis 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 psychosomatic 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 conferenceof 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 International 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 detection of textual cyberbullying." in The Social Mobile Web, 2011.
[12]. V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," Communications in Information Science and Management Engineering, 2012.
[13]. M. Dadvar, F. de Jong, R. Ordelman, and R. Trieschnigg, "Improved 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, "Improving 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 architectures," Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7, p. 43, 2012.
[17]. M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," arXiv preprint arXiv: 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," Proceedings 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 semantic 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 grounds? 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 children 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 Personalization. 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.