International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 07
June 2017

Semantic Enhanced Marginalized De-Noising Auto Encoder as a Learning Model for Cyber Bullying Detection

1.J.PRIYANKA 2. L.SRINIVASA RAO 3. DR. K NAGESWARARAO

[1]Pg Scholar, Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
jinugupriyanka@gmail.com
[2]Associate Professor , Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
srinu.pha4@gmail.com.

[3] Professor & HOD, Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
nageswararaokapu@yahoo.com.

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 ofsemantic 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.

1 INTRODUCTION

SOCIAL Media, as defined in a group of Internet- based 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 face-to-face 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 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 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. 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. 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. 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 vector machine for online harassment detection [10]. Dinakar et.al utilized label specific features to extend the general features, where the label specific features are learned by Linear Discriminative Analysis [11]. 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 [12]. Besides content-based information, Maral et.al proposed to apply users' information, such as gender and history messages, and context information as extra features . 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. 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 (mS- DA) 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 Au- toencoders (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. 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. In 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 ofour work can be summarized as follows:

* Our proposed Semantic-enhanced Marginalized S- tacked 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 re-construction 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.

2 RE LATED WORK

This work aims to learn a robust and discriminative text representation for cyberbullying detection. Text representation and automatic cyberbullying detection are both related to our work. In the following, we briefly review the previous work in these two areas.

2.1 Text Representation Learning

In text mining, information retrieval and natural language processing, effective numerical representation of linguistic units is a key issue. The Bag-of-words (BoW) model is the most classical text representation and the cornerstone of some states-of-arts models including Latent Semantic Analysis (LSA) and topic models . BoW model represents a document in a textual corpus using a vector of real numbers indicating the occurrence of words in the document. Although BoW model has proven to be efficient and effective, the representation is often very sparse. To address this problem, LSA applies Singular Value Decomposition (SVD) on the word-document matrix for BoW model to derive a low-rank approximation. Each new feature is a linear combination of all original features to alleviate the sparsity problem. Topic models, including Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation are also proposed. The basic idea behind topic models is that word choice in a document will be influenced by the topic of the document probabilistically. Topic models try to define the generation process of each word occurred in a document. Similar to the approaches aforementioned, our proposed approach takes the BoW

representation as the input. Hoever, our approach has some distinct merits. Firstly, the multilayers and non-linearity of our model can ensure a deep learning architecture for text representation, which has been proven to be effective for learning high-level features. Second, the applied dropout noise can make the learned representation more robust. Third, specific to cyberbullying detection, our method employs the semantic information, including bullying words and sparsity constraint imposed on mapping matrix in each layer and this will in turn produce more discriminative representation.

3 Cyberbullying Detection

With the increasing popularity of social media in recent years, cyberbullying has emerged as a serious problem afflicting children and young adults. Previous studies of cyberbullying focused on extensive surveys and its psychological effects on victims, and were mainly conducted by social scientists and psychologists. Although these efforts facilitate our understanding for 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 attracted 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 training. 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.

As an introductory work, they did not develop specialized models for cyberbullying detection. Yin et.al proposed to combine BoW features, sentiment feature and contextual features to train a classifier for detecting possible harassing posts. The introduction of the sentiment and contex-

tual features has been proven to be effective. Dinakar et.al used Linear Discriminative Analysis to learn label specific features and combine them with BoW features to train a

classifier [11]. 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. Although the incorporation of knowledge base can achieve a performance improvement, the construction of a complete and general one is labor-consuming. Nahar et.al proposed to scale bullying words by a factor of two in the original BoW features . The motivation behind this work is quit similar to that of our model to enhance bullying features. However, the scaling operation in is quite arbitrary. Ptaszynski et.al searched sophisticated patterns in a brute-force way. The weights for each extracted pattern need to be calculated based on annotated training corpus, and thus the performance may not be guaranteed if the training corpus has a limited size. Besides content-based information, Maral et.al also employ users' information, such as gender and history messages, and context information as extra features [13],

. Huang et.al also considered social network features to learn the features for cyberbullying detection. The shared deficiency among these for mentioned approaches is constructed text features are still from BoW representation, which has been criticized for its inherent over-sparsity and failure to capture semantic structure . Different from these approaches, our proposed model can learn robust features by reconstructing the original data from corrupted data and introduce semantic corruption noise and sparsity mapping matrix to explore the feature structure which are predictive of the existence of bullying so that the learned representation can be discriminative. Marginalized Denoising Auto-encoder In what follows, we describe our approach. The key idea is to marginalize out the noise of the corrupted inputs in the denoising auto-encoders. We start by describing the conventional denoising auto-encoders and introducing necessary notations. Afterwards, we present the detailed derivations of our approach. Our approach is general and flexible to handle various types of noise and loss functions for denoising. A few concrete examples with popular choices of noise and loss functions are included for illustration. We then analyze the properties of the proposed approach while drawing connections to existing works. 2.1. Denoising Auto-encoder (DAE) The Denoising Auto-Encoder (DAE)

is typically implemented as a one-hidden-layer neural network which is trained to reconstruct a data point x ∈ RD from its (partially) corrupted version x˜ (Vincent et al., 2008). The corrupted input x˜ is typically drawn from a conditional distribution p(x˜|x) — common corruption choices are additive Gaussian noise or multiplicative mask-out noise (where values are set to 0 with some probability q and kept unchanged with probability of 1 − q). The corrupted input x˜ is first mapped to a latent representation through the encoder (i.e., the nonlinear transformation between the input layer and the hidden layer). Let z=hθ(x˜) ∈ RDh denote the Dh-dimensional latent representation, collected at the outputs of the hidden layer. The code z is then decoded into the network output y = gθ(z) ∈ RD by the nonlinear mapping from the hidden layer to the output layer. Note that we follow the custom to have both mappings share the same parameter θ. For denoising, we desire y = g ◦ h(x˜) = fθ(x˜) to be as close as possible to the clean data x. To this end, we use a loss function `(x, y) to measure the reconstruction error. Given a dataset D = {x1, · · · , xn}, we optimize the parameter θ by corrupting each xi m-times, yielding x˜ 1 i , . . . , x˜m i , and minimize the averaged reconstruction loss 1 n Xn i=1 1 m Xm j=1 ` xi , fθ(x˜ j i)) . (1) Typical choices for the loss ` are the squared loss for realvalued inputs, or the cross-entropy loss for binary inputs

CONCLUSION

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 Term Reconstruction on Twitter datasets. Each Row Shows Specific Bullying Word, along with Top-4 Reconstructed Words (ranked with their frequency values from top to bottom) via

mSDA (left column) and smSDA (right column).

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 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 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, 199