*International Journal of Research in Advent Technology, Special Issue, March 2019*
*E-ISSN: 2321-9637*
*International Conference on Technological Emerging Challenges (ICTEC-2019)*
*Available online at www.ijrat.org*

# Recognition Of Cyberbulling Based On The Unconscious Code Of Disregarded Denoising Improved Semantic

[1]Matcha Deepthi Sri, [2]Manchala Pujitha, [3]Ganga Vasantha, [4]Medaboyina Nagarjuna Yadav
[5]Achukola Venkata Sai Krishna ,[6]Mr.B.Ramesh

**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 isrobust 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 developedvia 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 theword 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 out perform other baseline text representation learningmethods.

## 1. INTRODUCTION

SOCIAL Media, as defined in a group of Internet-based applications that build on the ideological andtechnological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.‘’ Via socialmedia, people can enjoy enormous information, convenientcommunication experience and so on. However, social media may have some side effects such as cyberbullying, whichmay have negative impacts on the life of people, especially children and teenagers. Cyberbullying can be defined as aggressive, intentionalactions performed by an individual or a group of people viadigital communication methods such as sending messagesand posting comments against a victim. Different from traditional bullying that usually occurs at school during face-to-face communication, cyberbullying on social media cantake place anywhere at any time. For bullies, they are freeto hurt their peers’ feelings because they do not need to facesomeone and can hide behind the Internet. For victims, theyare easily exposed to harassment since all of us, especiallyyouth, are constantly connected to Internet or social media. As reported in cyberbullying victimization rate rangesfrom 10%to 40%. In the United States, approximately 43%of teenagers were ever bullied on social media . Thesame as traditional bullying, cyberbullying has negative,insidious and sweeping impacts on children .The outcomes for victims under cyberbullying may evenbe tragic such as the occurrence of self-injurious behavior or suicides.One way to address the cyberbullying problem is toautomatically detect and promptly report bullying messageso that proper measures can be taken to prevent possible tragedies. Previous works on computational studiesof bullying have shown that natural language processingand machine learning are powerful tools to study bullying. Cyberbullying detection can be formulated as asupervised learning problem. A classifier is first trained ona cyberbullying corpus labeled by humans, and the learnedclassifier is then used to recognize a bullying message.Three kinds of information including text, user demography,and social network features are often used in cyberbullyingdetection. Since the text content is the most reliable, ourwork here focuses on text-based cyberbullying detection.

In the text-based cyberbullying detection, the first andalso critical step is the numerical representation learningfor text messages. In fact, representation learning of text isextensively studied in text mining, information retrieval andnatural language processing (NLP). Bag-of-words (BoW)model is one commonly used model that each dimensioncorresponds to a term. Latent Semantic Analysis (LSA)and topic modelsare another popular text representationmodels, which are both based on BoW models. By mappingtext units into fixed-length vectors, the learned representation can be further processed for numerous languageprocessing 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 messageson social media are often very short and contain a lot ofinformal language and misspellings, robust representationsfor these messages are required to reduce their ambiguity. Even worse, the lack of sufficient high-quality trainingdata, 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 froma third view due to its intrinsic ambiguities. Thirdly, dueto protection of Internet users and privacy issues, only asmall portion of messages are left on the Internet, and mostbullying posts are deleted. As a result, the trained classifiermay not generalize well on testing messages that containnonactivated but discriminative features. The goal of thispresent 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 theseproblems by incorporating expert knowledge into featurelearning. 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]. Dinakaret.al utilized label specific features to extend the generalfeatures, where the label specific features are learned byLinear Discriminative Analysis [11]. In addition, commonsense knowledge was also applied. Nahar et.al presented a weighted TF-IDF scheme via scaling bullying-like featuresby a factor of two [12]. Besides content-based information, Maral et.al proposed to apply users' information, such asgender and history messages, and context information as extra features . But a major limitation of theseapproaches is that the learned feature space still relies onthe BoW assumption and may not be robust. In addition,the performance of these approaches rely on the qualityof hand-crafted features, which require extensive domainknowledge.In this paper, we investigate one deep learning methodnamed stacked denoisingautoencoder (SDA) . SDAstacks several denoisingautoencoders and concatenates theoutput of each layer as the learned representation. Eachdenoisingautoencoder in SDA is trained to recover theinput data from a corrupted version of it. The input iscorrupted by randomly setting some of the input to zero,which is called dropout noise. This denoising process helpsthe autoencoders to learn robust representation. In addition,each autoencoder layer is intended to learn an increasinglyabstract representation of the input . In this paper, wedevelop a new text representation model based on a variantof SDA: marginalized stacked denoisingautoencoders (mS- DA) which adopts linear instead of nonlinear projectionto accelerate training and marginalizes infinite noise distribution in order to learn more robust representations. Weutilize semantic information to expand mSDA and developSemantic-enhanced Marginalized Stacked Denoising Au-toencoders (smSDA). The semantic information consists ofbullying words. An automatic extraction of bullying wordsbased on word embeddings is proposed so that the involvedhuman labor can be reduced. During training of smSDA, weattempt to reconstruct bullying features from other normalwords by discovering the latent structure, i.e. correlation,

between bullying and normal words. The intuition behindthis idea is that some bullying messages do not containbullying words. The correlation information discovered bysmSDA helps to reconstruct bullying features from normalwords, and this in turn facilitates detection of bullyingmessages without containing bullying words. For example,there is a strong correlation between bullying word fuck andnormal 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 mayhelp to reconstruct the bullying features from normal onesso that the bullying message can be detected. It shouldbe noted that introducing dropout noise has the effects ofenlarging the size of the dataset, including training datasize, which helps alleviate the data sparsity problem. Inaddition, L1 regularization of the projection matrix is addedto the objective function of each autoencoder layer in ourmodel to enforce the sparstiy of projection matrix, and thisin turn facilitates the discovery of the most relevant termsfor reconstructing bullying terms. The main contributions ofour work can be summarized as follows:

* Our proposed Semantic-enhanced Marginalized S-tacked DenoisingAutoencoder is able to learn robust features from BoW representation in an efficient and effective way. These robust features arelearned by reconstructing original input from corrupted (i.e., missing) ones. The new feature spacecan 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 semanticdropout noises and imposing sparsity constraintson mapping matrix. In our framework, high-qualitysemantic information, i.e., bullying words, can beextracted automatically through word embeddings.Finally, these specialized modifications make thenew feature space more discriminative and this inturn facilitates bullying detection.

* Comprehensive experiments on real-data sets haveverified the performance of our proposed model.

## 2. RELATED WORK

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

## 3. TEXT REPRESENTATIONLEARNING

In text mining, information retrieval and natural languageprocessing, effective numerical representation of linguisticunits is a key issue. The Bag-of-words (BoW) model isthe most classical text representation and the cornerstoneof some states-of-arts models including Latent SemanticAnalysis (LSA) and topic models . BoW modelrepresents a document in a textual corpus using a vectorof real numbers indicating the occurrence of words in thedocument. Although BoW model has proven to be efficientand effective, the representation is often very sparse. Toaddress this problem, LSA applies Singular Value Decomposition (SVD) on the word-document matrix for BoW modelto derive a low-rank approximation. Each new feature is alinear combination of all original features to alleviate thesparsity problem. Topic models, including Probabilistic Latent Semantic Analysis and Latent DirichletAllocation are also proposed. The basic idea behind topic modelsis that word choice in a document will be influenced by thetopic of the document probabilistically. Topic models try todefine the generation process of each word occurred in adocument.Similar to the approaches aforementioned, our proposedapproach 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 deeplearning architecture for text representation, which has beenproven to be effective for learning high-level features.Second, the applied dropout noise can make the learnedrepresentation more robust. Third, specific to cyberbullyingdetection, our method employs the semantic information,including bullying words and sparsity constraint imposedon mapping matrix in each layer and this will in turn produce more discriminative representation.

## 4. CYBERBULLYING DETECTION

With the increasing popularity of social media in recentyears, cyberbullying has emerged as a serious problemafflicting children and young adults. Previous studies ofcyberbullying focused on extensive surveys and its psychological effects on victims, and were mainly conductedby social scientists and psychologists.Although these efforts facilitate our understanding for cyberbullying, the psychological science approach based onpersonal surveys is very time-consuming and may notbe suitable for automatic detection of cyberbullying. Sincemachine learning is gaining increased popularity in recentyears, the computational study of cyberbullying has attracted the interest of researchers. Several research areasincluding topic detection and affective analysis are closelyrelated to cyberbullying detection. Owing to their efforts,automatic cyberbullying detection is becoming possible. Inmachine learning-based cyberbullying detection, there aretwo issues:

1) text representation learning to transform eachpost/message into a numerical vector and

2) classifier training. Xu et.al presented several off-the-shelf NLP solutionsincluding BoW models, LSA and LDA for representationlearning to capture bullying signals in social media.

As an introductory work, they did not develop specializedmodels for cyberbullying detection. Yin et.al proposed tocombine BoW features, sentiment feature and contextualfeatures 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.alused Linear Discriminative Analysis to learn label specificfeatures and combine them with BoW features to train aclassifier [11]. The performance of label-specific featureslargely depends on the size of training corpus. In addition,
they need to construct a bullyspace knowledge base to boostthe performance of natural language processing methods.Although the incorporation of knowledge base can achievea performance improvement, the construction of a completeand general one is labor-consuming. Nahar et.al proposed toscale bullying words by a factor of two in the original BoWfeatures . The motivation behind this work is quit similarto that of our model to enhance bullying features. However, the scaling operation in is quite arbitrary. Ptaszynskiet.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 alimited size. Besides content-based information, Maral et.alalso employ users' information, such as gender and historymessages, and context information as extra features [13],
. Huang et.al also considered social network featuresto learn the features for cyberbullying detection. Theshared deficiency among these formentioned approaches isconstructed text features are still from BoW representation,which has been criticized for its inherent over-sparsity andfailure to

capture semantic structure . Different from these approaches, our proposed model can learnrobust features by reconstructing the original data fromcorrupted data and introduce semantic corruption noise andsparsity mapping matrix to explore the feature structurewhich are predictive of the existence of bullying so that thelearned 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˜mi , and minimize the averaged reconstruction loss 1 n Xni=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 .

## 5. CONCLUSION

This paper addresses the text-based cyberbullying detectionproblem, where robust and discriminative representationsof messages are critical for an effective detection system.By designing semantic dropout noise and enforcing sparsity, we have developed semantic-enhanced marginalizeddenoisingautoencoder as a specialized representation learning model for cyberbullying detection. In addition, wordembeddings have been used to automatically expand andrefine bullying word lists that is initialized by domainknowledge. The performance of our approaches has beenexperimentally verified through two cyberbullying corporafrom social medias: Twitter and MySpace. As a next stepwe are planning to further improve the robustness of theTerm Reconstruction on Twitter datasets. Each Row Shows SpecificBullying Word, along with Top-4 Reconstructed Words (ranked withtheir frequency values from top to bottom) via mSDA (left column) andsmSDA (right column).

## REFERENCES

[1] A. M. Kaplan and M. Haenlein, “Users of the world, unite! Thechallenges 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-sexualaggression over time and linkages to nontechnology aggression,”National Summit on Interpersonal Violence and Abuse Across theLifespan: Forging a Shared Agenda, 2010.

[4] B. K. Biggs, J. M. Nelson, and M. L. Sampilo, “Peer relations inthe 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 ofbullying 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 miningand 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 frombullying traces in social media,” in Proceedings of the 2012 conferenceof the North American chapter of the association for computationallinguistics: Human language

technologies. Association for Computational Linguistics, 2012, pp. 656–666.
[9] Q. Huang, V. K. Singh, and P. K. Atrey, “Cyber bullying detectionusing 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, andL. Edwards, “Detection of harassment on web 2.0,” Proceedingsof 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 andManagement Engineering, 2012.
[13] M. Dadvar, F. de Jong, R. Ordelman, and R. Trieschnigg, “Im-proved cyberbullying detection using gender information,” inProceedings 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 inInformation Retrieval. Springer, 2013, pp. 693–696.
[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol,“Stacked denoisingautoencoders: Learning useful representationsin a deep network with a local denoising criterion,” The Journal ofMachine Learning Research, vol. 11, pp. 3371–3408, 2010.
[16] P. Baldi, “Autoencoders, unsupervised learning, and deep archi-tectures,” Unsupervised and Transfer Learning Challenges in MachineLearning, 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 tolatent semantic analysis,” Discourse processes, vol. 25, no. 2-3, pp.259–284, 199