Cyberblustery Detection Using Machine Learning Techniques

S. Priyadharshini
Department of Computer Science and Engineering
TRP Engineering College,
Tiruchirappalli, India.

K. Simladevi
Department of Computer Science and Engineering
TRP Engineering College,
Tiruchirappalli, India.

R. Sathya
Department of Computer Science and Engineering
TRP Engineering College,
Tiruchirappalli, India.

***Abstract*—As more people turn to the Internet for school, work, and social use, so too do more people turn to the Internet to take out their frustrations and aggression. One form of cyber aggression has been gaining the attention of both researchers and the public in recent years: cyber-bullying. Cyber-bullying is typically defined as aggression that is intentionally and repeatedly carried out in an electronic context (e.g., e-mail, blogs, instant messages, text messages) explicitly corrupted. In this paper, we present the marginalized Denoising Auto-encoder (mDAE), which (approximately) marginalizes out the corruption during training. Effectively, the mDAE takes into account infinitely many corrupted copies of the training data in every epoch, and therefore is able to match or outperform the DAE with much fewer training epochs. Our proposed algorithm and show that it can be understood as a classic auto-encoder with a special form of regularization. Then implement the framework to analyze the approach cyber bullying words to automatically change the words and recommend the possible words. The experimental results shows that the proposed approach can be provide improved accuracy than the existing approaches.**

Keywords*— *Cyberblustery detection, Textmining, Word-replacement, Stacked denoising auto encoder.

I. INTRODUCTION

Social media, as defined as a group of internet based applications that build on the ideological and technological foundations of web 2.0,and that allow the creation 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 in 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 [4], [5], [6]. 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 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 [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, labelling data is labour 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 non activated 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, sentiment 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. 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 . 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 auto-encoder (SDA)]. SDA stacks several denoising auto-encoders and concatenates the output of each layer as the learned representation. Each denoising auto-encoder 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 auto-encoders to learn robust representation. In addition, each auto-encoder 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 auto-encoders (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 Auto-encoders (smSDA). The semantic information consists of bullying words. An automatic extraction of bullying words based on word embedding's is proposed so that the involved human labour 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 auto-encoder layer in our model to enforce the sparsity of projection matrix, and this in turn facilitates the discovery of the most relevant terms for reconstructing bullying terms.

II. EXISTING METHOD

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. 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. On training the corpus they need to construct a bully space 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. 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 .

Huang et.al also considered social network features to learn the features for cyberbullying detection. The shared deficiency among these fore 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 Stacked Denoising Auto-encoder

In this model, denoising auto-encoder attempts to reconstruct original data using the corrupted data via a linear projection. The advantage of corrupting the original input in mSDA can be explained by feature co-occurrence statistics. The co-occurrence information is able to derive a robust feature representation under an unsupervised learning framework, and this also motivates other state-of-the-art next feature learning methods such as Latent Semantic Analysis and topic models, a denoising auto-encoder is trained to reconstruct these removed features values from the rest uncorrupted ones. It is shown that the learned representation is robust and can be regarded as s high level concept feature since the correlation information is invariant to domain-specific vocabularies. The extension of mSDA include semantic dropout noise and sparse mapping constraints..

Semantic Dropout noice

The correlation explored by the auto encoder structure enables the subsequent classifier to learn the discriminative word and improve the classification performance. Tn addition, the semantic dropout noise exploits the correlation between the bullying features and normal features better and hence facilitates cyberbullying detection.

Construction of bullying feature set

The bullying features plays an important role and should be chosen properly in terms of layers. 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 sand regard the intersections as bullying features. However, it is possible that the expert knowledge is limited and does not reflect the current usage and style of cyber language. Therefore, we expand the list of pre-defined insulting words based on word embeddings. Word embeddings use real-valued and low-dimensional vectors to represent semantics of words. The well-trained word embeddings lie in a vector space where similar words are placed close to each other. Considering the interent messages are our interested corpus , we utilize a well trained word to vector model on a large scale. It is observed that curse words form distinct clusters ,which are also far away from normal words. Even insulting words are located at different regions due to different word usages and insulting expressions. In addition ,the cosine similarity between word embeddings is able to quantify the semantic similarity between words. For each insulting seed, similar words are extracted if their cosine similarities with insult seed exceed a predefined threshold .For bigram, W_1W_r,we simply use an additive model to derive the corresponding embedding as follows:

$$v(W_1W_r) = v(W_1) + v(W_r)$$

We perform feature selection using Fisher score to select “bullying” features. Fisher score is an univariate metric reflecting the discriminative power of a feature.

Sparsity Constraints

The In mSDA, the mapping matrix is learned to reconstruct removed features from other uncorrupted features and hence is able to capture the feature correlation information. Here, we inject the sparsity constraints on the mapping weights so that each row has a small number of nonzero elements. This sparsity constraint is quite intuitive because one word is only related to a small portion of vocabulary instead of the whole vocabulary. In our proposed smSDA, the sparsity constraint is realized by the incorporation of regularization term into the objective function as in the lasso problem.

Merits of smSDA

The Some important merits of our proposed approach are summarized as follows:

- Most cyberbullying detection methods rely on the BoW model. Due to the sparsity problems of both data and features, the classifier may not be trained very well. Stacked densoing auto-encoder (SDA), as an unsupervised representation learning method, is able to learn a robust feature space. In SDA, the feature correlation is explored by the reconstruction of corrupted data. The learned robust feature representation can then boost the training of classifier and finally improve the classification accuracy. In addition, the corruption of data in SDA actually generates artificial data to expand data size, which alleviate the small size problem of training data.
- For cyberbullying problem, we design semantic dropout noise to emphasize bullying features in the new feature space, and the yielded new representation is thus more discriminative for cyberbullying detection.
- The sparsity constraint is injected into the solution of mapping matrix W for each layer, considering each word is only correlated to a small portion of the whole vocabulary. We formulate the solution

for the mapping weights W, as an Iterated Ridge Regression problem, in which the semantic dropout noise distribution can be easily marginalized to ensure the efficient training of our proposed smSDA.

- 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.

III. PROPOSED METHOD

Social network refers to the application of interaction among people in which they create, share and exchange information and ideas in virtual communities and networks. Firstly, the users will start to interact by sending them some messages. Those messages and comments will be sent and stored in server database. The server will represents the text by splitting into phrases. From the represented phrases, the keywords are picked and check whether it is bullying word or not. If not, they will be sent. If the words or texts are considered as the bullying words, the system will seek for the next phase of finding semantic meaning of those identified bullying words. The semantic meaning lies in the word net library which is connected to the system. The appropriate words with related meaning is found out for the reconstruction process. After finding out ,they will be replaced instead of those bullying words. Then the user can communicate with another user through filtered GUI.

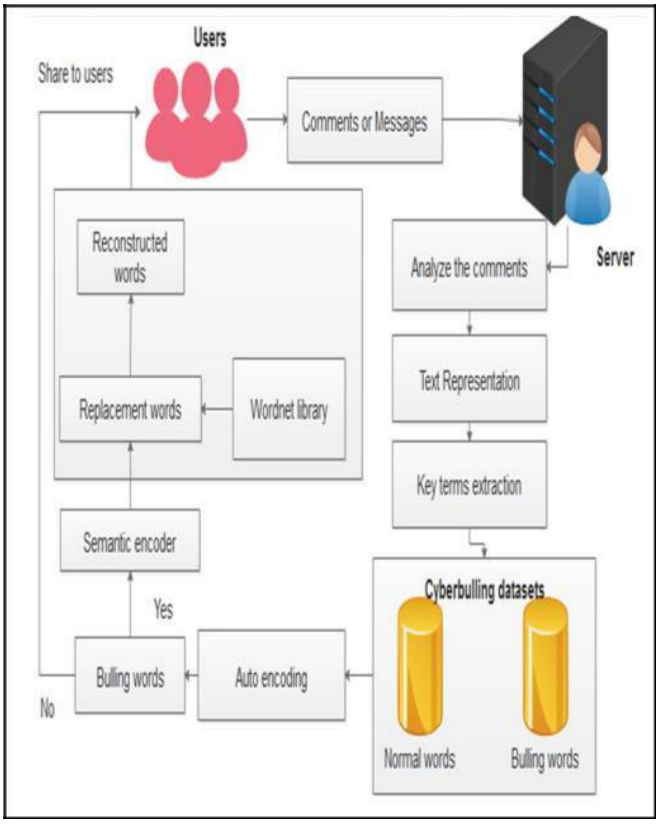

Fig.1 : Architecture of cyberblustery replacement

TABLE I

Bullying words	Reconstructed words for	
	mSDA	*SmSDA*
Bitch	@USER Shut Friend Tell	@USER HTTPLINK Fuck up Shut
Fucking	Because Friend Off Gets	Off Pissed Shit Of
Shit	Some Big With Lol	Abuse This shit Shit lol Big

Fig. 2: Term Reconstruction For Facebook Data sets

Regularized auto-encoders are important building blocks for learning deep and rich representations of data. The standard approach of denoising auto-encoder incorporates regularization via learning reconstruction from partially corrupted samples. While effective, this is often a computationally intensive and lengthy process. The mDAE overcomes the limitation by marginalizing the corruption process, effectively learning from infinitely many corrupted samples. At the core of our approach is to approximate the expected loss function with its Taylor expansion. Our analysis yields a regularization term that takes into consideration both the reconstruction function's sensitivity to the hidden representations and the hidden representation's sensitivity to the inputs. The main objective of the paper is to partition abusive messages from big data streaming with encoder method on top of word probabilistic on each document for determining the similarity sentences score based on the improving accuracy and computation time. The paper proposes a novel method which can generate a predictive model from large volume of data sets for supporting the analysis services on business.

REFERENCES

[1] Zhao,Kezhi Mao,"Cyber bullying detection based on semantic-enhanced marginalized denoising auto-encoder",Vol 8,n0.3,July September 2017.

[2] 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 J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] P.Baldi,"Autoencoders, unsupervised learning and deep architectures", Unsupervised and Transfer Learning Challenges in Machine Learning,Volume 7,p.43,2012.

[4] T.Mikolov,K.Chen, G.S.Corrado and J.Dean,"Efficient estimation of word representation in vector space",arXiv preprint arXiv:1301.3781,2013.

[5] T,Mikolov,I.Sutskever,K.chen,G.S.Corradoand J.Dean,"Distributed representations of words and phrases and their compositionality",inAdvances in neural information 2001.

[6] A.Kontostathis,L.Edwards and A.Leatherman,"Textmining and cybercrime",Text mining:Applications and theory.John wiley and Sons,Ltd,Chichester,UK,2010.
[7] J.Fan and R.Li,"Variable selection via nonconcave penalized likelihood and its oracle properties,"Journal of the American statistical Association,vol.96,no.456,pp.1348-1360.
[8] J.Sui,"Understanding and fighting bullying with machine learning",Ph.D. dissertation, The UNIVERSITY OF WISCONSIN-MADISON-2015.
[9] T.K.Landauer,P.W.Foltz and D.Laham,"An introduction to latent semantic analysis",Discourse process.vol.25, no. 2-3,pp.259-284,1998.
[10] K.Dinakar,R.Reichart and H.Lieberman,"Modeling the detection of textual Cyberbullying."in The social Mobile Web,2011.