Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

Rui Zhao and Kezhi Mao

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.

Index Terms—Cyberbullying Detection, Text Mining, Representation Learning, Stacked Denoising Autoencoders, Word Embedding

1 Introduction

SOCIAL Media, as defined in [1], is ''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 [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 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 [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

- *R. Zhao and K. Mao are with with the School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798.*
E-mail: rzhao001,ekzmao@ntu.edu.sg

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, 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 [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 [13], [14]. 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) [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 [16]. In this paper, we develop a new text representation model based on a variant of SDA: marginalized stacked denoising autoencoders (mSDA) [17], 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. 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 of our work can be summarized as follows:

* 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.

This paper is organized as follows. In Section 2, some related work is introduced. The proposed Semantic-enhanced Marginalized Stacked Denoising Auto-encoder for cyberbullying detection is presented in Section 3. In Section 4, experimental results on several collections of cyberbullying data are illustrated. Finally, concluding remarks are provided in Section 5.

2 Related 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) [18] and topic models [19], [20]. 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 [21] and Latent Dirichlet Allocation [20], 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. However, our approach has some distinct merits. Firstly, the multi-layers 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 [22]. 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.

2.2 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 [6], [23], [24], [25]. Although these efforts facilitate our understanding for cyberbullying, 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 [8]. 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 [10]. The introduction of the sentiment and contextual 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 [12]. The motivation behind this work is quit similar to that of our model to enhance bullying features. However, the scaling operation in [12] is quite arbitrary. Ptaszynski et.al searched sophisticated patterns in a brute-force way [26]. 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], [14]. Huang et.al also considered social network features to learn the features for cyberbullying detection [9]. The shared deficiency among these forementioned 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 [18], [19], [20]. 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.

3 Semantic-Enhanced Marginalized Stacked Denoising Auto-encoder

We first introduce notations used in our paper. Let $D = \{w_1, \ldots, w_d\}$ be the dictionary covering all the words existing in the text corpus. We represent each message using a BoW vector $\mathbf{x} \in \mathbb{R}^d$. Then, the whole corpus can be denoted as a matrix: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where n is the number of available posts.

We next briefly review the marginalized stacked denoising auto-encoder and present our proposed Semantic-enhanced Marginalized Stacked Denoising Auto-Encoder.

3.1 Marginalized Stacked Denoising Auto-encoder

Chen et.al proposed a modified version of Stacked Denoising Auto-encoder that employs a linear instead of a non-linear projection so as to obtain a closed-form solution [17]. The basic idea behind denoising auto-encoder is to reconstruct the original input from a corrupted one $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ with the goal of obtaining robust representation.

Marginalized Denoising Auto-encoder: In this model, denoising auto-encoder attempts to reconstruct original data using the corrupted data via a linear projection. The projection matrix can be learned as:

$$\mathbf{W} = \underset{\mathbf{W}}{\mathrm{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{W}\tilde{\mathbf{x}}_i\|^2 \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{d \times d}$. For simplicity, we can write Eq. (1) in matrix form as:

$$\mathbf{W} = \underset{\mathbf{W}}{\mathrm{argmin}} \frac{1}{2n} tr \left[(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}})^{\mathrm{T}} (\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}}) \right] \quad (2)$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]$ is the corrupted version of $\mathbf{X}$. It is easily shown that Eq. (2) is an ordinary least square problem having a closed-form solution:

$$\mathbf{W} = \mathbf{P}\mathbf{Q}^{-1} \tag{3}$$

where $\mathbf{P} = \mathbf{X}\tilde{\mathbf{X}}^{\mathrm{T}}$ and $\mathbf{Q} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}$. In fact, this corruption can be marginalized over the noise distribution [17]. The more corruptions we take in the denoising auto-encoder, the more robust transformation can be learned. Therefore, the best choice is using infinite versions of corrupted data. If the data corpus is corrupted infinite times, the matrix P and Q are converged to their corresponding expectation, and Eq. (3) can be formulated as:

$$\mathbf{W} = E\left[\mathbf{P}\right] E\left[\mathbf{Q}\right]^{-1} \tag{4}$$

where $E\left[\mathbf{P}\right] = \sum_{i=1}^{n} E\left[\mathbf{x}_i\tilde{\mathbf{x}}_i^{\mathrm{T}}\right]$ and $E\left[\mathbf{Q}\right] = \sum_{i=1}^{n} E\left[\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^{\mathrm{T}}\right]$. These expected matrices can be computed based on noise distribution. In [17], dropout noise is adopted to corrupt data samples by setting a feature to zero with a probability p. Assuming the scatter matrix of the original data samples is denoted as $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, the expected matrices can be computed as:

$$E\left[\mathbf{Q}\right]_{i,j} = \begin{cases} (1-p)^2\mathbf{S}_{i,j} & \text{if } i \neq j, \\ (1-p)\mathbf{S}_{i,j} & \text{if } i = j. \end{cases} \tag{5}$$

and

$$E\left[\mathbf{P}\right]_{i,j} = (1-p)\mathbf{S}_{i,j} \tag{6}$$

where i and j denotes the indices of features. It can be seen that it is very efficient to compute $\mathbf{W}$ by marginalizing dropout noise in denoising auto-encoder. After the mapping weights $\mathbf{W}$ are computed, a nonlinear squashing function, such as a hyperbolic tangent function, can be applied to derive the output of the marginalized denoising auto-encoder:

$$\mathbf{H} = \tanh(\mathbf{W}\mathbf{X}) \tag{7}$$

Stacking Structure: Chen et.al [17] also proposed to apply stacking structures on marginalized denoising autoencoder, in which the output of the $(k-1)^{th}$ layer is fed as the input into the k^{th} layer. If we define the output of the k^{th} mDA as $\mathbf{H}_k$ and the original input as $\mathbf{H}_0$ respectively, the mapping between two consecutive layers is given as:

$$\mathbf{H}_k = \tanh(\mathbf{W}_k\mathbf{H}_{k-1}) \tag{8}$$

where $\mathbf{W}_k$ denotes the mapping in k^{th} layer. The model training can be done greedily layer by layer. This means that the mapping weights $\mathbf{W}_k$ is learned in a closed-form to reconstruct the output of $(k-1)^{th}$mDA layer from its marginalized corruptions, as shown in Eq. (4). If the number of layers is set to L, the final representation for input data $\mathbf{X}$ is the concatenation of the uncorrupted original input and outputs of all layers as follows:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_L \end{bmatrix} \tag{9}$$

where $\mathbf{Z} \in \mathbb{R}^{d(L+1)\times n}$. Each column of $\mathbf{Z}$ represents the final representation of each individual data sample.

3.2 Semantic Enhancement for mSDA

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 text feature learning methods such as Latent Semantic Analysis and topic models [18], [20]. As shown in Figure 1. (a), a denoising auto-encoder is trained to reconstruct these removed features values from the rest uncorrupted ones. Thus, the learned mapping matrix W is able to capture correlation between these removed features and other features. It is shown that the learned representation is robust and can be regarded as a high level concept feature since the correlation information is invariant to domain-specific vocabularies. We next describe how to extend mSDA for cyberbullying detection. The major modifications include semantic droupout noise and sparse mapping constraints.

3.2.1 Semantic Dropout Noise

The dropout noise adopted in mSDA is an uniform distribution, where each feature has the same probability to be removed. In cyberbullying detection, most bullying posts contain bullying words such as profanity words and foul languages. These bullying words are very predictive of the existence of cyberbullying. However, a direct use of these bullying features may not achieve good performance because these words only account for a small portion of the whole vocabulary and these vulgar words are only one kind of discriminative features for bullying [10], [26]. In other way, we can explore these cyberbullying words by using a different dropout noise that features corresponding to bullying words have a larger probability of corruption than other features. The imposed large probability on bullying words emphasizes the correlation between bullying features and normal ones. This kind of dropout noise can be denoted as semantic dropout noise, because semantic information is used to design dropout structure.

As shown in Figure 1. (b), the correlation between features can enable other normal words to predict bullying labels. Considering a simple but intuitive example, "Leave him alone, he is just a chink"[1], which is obviously a bullying message. However, the classifier will set the weight of the discriminative word "chink" to zero, if the small sized training corpus does not cover it. Our proposed smSDA can deal with the problem by learning a robust feature representation, which is a high level concept representation. In the learned representation, the word "chink" are reconstructed by context words co-occurring with the specific word ("chink") and the context words may be shared by other bullying words contained in training corpus. Therefore, the correlation explored by this auto-encoder structure enables the subsequent classifier to learn the discriminative word and improve the classification performance. In addition, the semantic dropout noise exploits the correlation between

[1] *"Chink (also chinki, chinky, chinkie) is an English ethnic slur usually referring to a person of Chinese or East Asian ethnicity"* from Wikipedia

bullying features and normal features better and hence, facilitates cyberbullying detection.

Due to the introduced semantic dropout noise, the expected matrices: $E\left[\mathbf{P}\right]$ and $E\left[\mathbf{Q}\right]$ will be computed slightly different from Eqs. (5) and (6). Assuming we have an available bullying words list and the corresponding features set $\mathbb{Z}_b$, the semantic dropout noise can be described as the following probability density function (PDF):

$$PDF = \begin{cases} p(\tilde{x}_d = 0) = p_n & \text{if } d \notin \mathbb{Z}_b, \\ p(\tilde{x}_d = x_d) = 1 - p_n & \text{if } d \notin \mathbb{Z}_b, \\ p(\tilde{x}_d = 0) = p_b & \text{if } d \in \mathbb{Z}_b, \\ p(\tilde{x}_d = x_d) = 1 - p_b & \text{if } d \in \mathbb{Z}_b, \end{cases} \tag{10}$$

where d denotes the feature set. Then these two marginalized matrices can be computed as:

$$E\left[\mathbf{Q}\right]_{i,j} = \begin{cases} (1-p_n)\mathbf{S}_{i,j} & \text{if } i = j \,\&\, i \notin \mathbb{Z}_b, \\ (1-p_n)^2\mathbf{S}_{i,j} & \text{if } i \neq j \,\&\, \{i,j\} \cap \mathbb{Z}_b = \varnothing, \\ (1-p_b)(1-p_n)\mathbf{S}_{i,j} & \text{if } \{i,j\} \notin \mathbb{Z}_b \,\&\, \{i,j\} \cap \mathbb{Z}_b \neq \varnothing, \\ (1-p_b)^2\mathbf{S}_{i,j} & \text{if } i \neq j \,\&\, \{i,j\} \in \mathbb{Z}_b, \\ (1-p_b)\mathbf{S}_{i,j} & \text{if } i = j \,\&\, i \in \mathbb{Z}_b. \end{cases} \tag{11}$$

and

$$E\left[\mathbf{P}\right]_{i,j} = \begin{cases} (1-p_n)\mathbf{S}_{i,j} & \text{if } j \cap \mathbb{Z}_b = \varnothing, \\ (1-p_b)\mathbf{S}_{i,j} & \text{if } j \cap \mathbb{Z}_b \neq \varnothing. \end{cases} \tag{12}$$

where p_b and p_n are the probabilities of bullying features and normal features to be set to zero respectively, and $p_b > p_n$. Here, p_b and p_n are both tunable hyperparameters for our proposed smSDA.

Unbiased Semantic Dropout Noise As shown in Eq . (10), the corrupted data is biased, i.e., $E\left[\mathbf{X}\right] \neq E\left[\tilde{\mathbf{X}}\right]$. Here, we modified Eq. (10) to achieve an unbiased noise as follows:

$$PDF^{unbiased} = \begin{cases} p(\tilde{x}_d = 0) = p_n & \text{if } d \notin \mathbb{Z}_b, \\ p(\tilde{x}_d = \frac{x_d}{1-p_n}) = 1 - p_n & \text{if } d \notin \mathbb{Z}_b, \\ p(\tilde{x}_d = 0) = p_b & \text{if } d \in \mathbb{Z}_b, \\ p(\tilde{x}_d = \frac{x_d}{1-p_b}) = 1 - p_b & \text{if } d \in \mathbb{Z}_b, \end{cases} \tag{13}$$

It can be easily shown that under such a noise distribution, the corrupted data is unbiased now. These two marginalized matrices are re-formulated as:

$$E\left[\mathbf{Q}\right]_{i,j}^{unbiased} = \begin{cases} \frac{1}{1-p_n}\mathbf{S}_{i,j} & \text{if } i = j \,\&\, i \notin \mathbb{Z}_b, \\ \frac{1}{1-p_b}\mathbf{S}_{i,j} & \text{if } i = j \,\&\, i \in \mathbb{Z}_b, \\ \mathbf{S}_{i,j} & \text{if } i \neq j. \end{cases} \tag{14}$$

and

$$E\left[\mathbf{P}\right]_{i,j}^{unbiased} = \mathbf{S}_{i,j} \tag{15}$$

These two computed matrices will then be used to learn the mapping in each layer in our proposed smSDA.

3.2.2 Sparsity Constraints

In mSDA, the mapping matrix $\mathbf{W}$ 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 $\mathbf{W}$ 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 L1 regularization term into the objective function as in the lasso problem [27]. The optimization function for each layer in smSDA is given as follows:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{2n} tr\left[(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}})^{\mathrm{T}}(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}})\right] + \lambda\|\mathbf{W}\|_1 \tag{16}$$

where λ is a regularization parameter that controls the sparsity of $\mathbf{W}$. The larger the λ is, the sparser the mapping matrix $\mathbf{W}$ is. The solution to Eq. (16) is a very mature math problem: sparse least squares optimization, which has several effective and efficient computation methods [28], [29], [30]. Here, we adopt a method called Iterated Ridge Regression, which has been proven to be very efficient [30]. The method firstly introduces an approximation:

$$\|\mathbf{w}_i\|_1 \approx \frac{\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i}{\|\mathbf{w}_i\|_1} \tag{17}$$

where $\mathbf{w}_i$ denotes the i-th row in the whole matrix $\mathbf{W}$. By substituting this approximation Eq. (17) into the objective function Eq. (16), we yield an formulation similar to a Ridge Regression Problem [31], and the iteration steps to solve $\mathbf{W}$ is given as:

$$\mathbf{W}_k = \tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{X}\left[\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} + \lambda\,\mathrm{diag}(|\mathbf{W}_{k-1}|)^{-1}\right]^{-1} \tag{18}$$

where diag denotes the diagonal elements of a matrix, $\mathbf{W}_k$ and $\mathbf{W}_{k-1}$ denote the current step and the previous step estimations for mapping matrix $\mathbf{W}$, respectively. It is clear that the Eq. (18) can be easily formulated when the noise distribution is marginalized. Similar to Eq. (4), Eq. (18) can be written as:

$$\mathbf{W}_k = E\left[\mathbf{P}\right]\left[E\left[\mathbf{Q}\right] + \lambda\,\mathrm{diag}(|\mathbf{W}_{k-1}|)^{-1}\right]^{-1} \tag{19}$$

To speed up the convergence process, the initialization for $\mathbf{W}$ can be set to the L2 penalized solution for Eq. (2) as follows:

$$\mathbf{W}_0 = E\left[\mathbf{P}\right]\left[E\left[\mathbf{Q}\right] + \lambda\mathbf{I}\right]^{-1} \tag{20}$$

where $\mathbf{I}$ is an identify matrix. It can be shown that this iteration procedure can also marginalize the noise distribution easily, which can ensure an efficient and stable mapping learning.

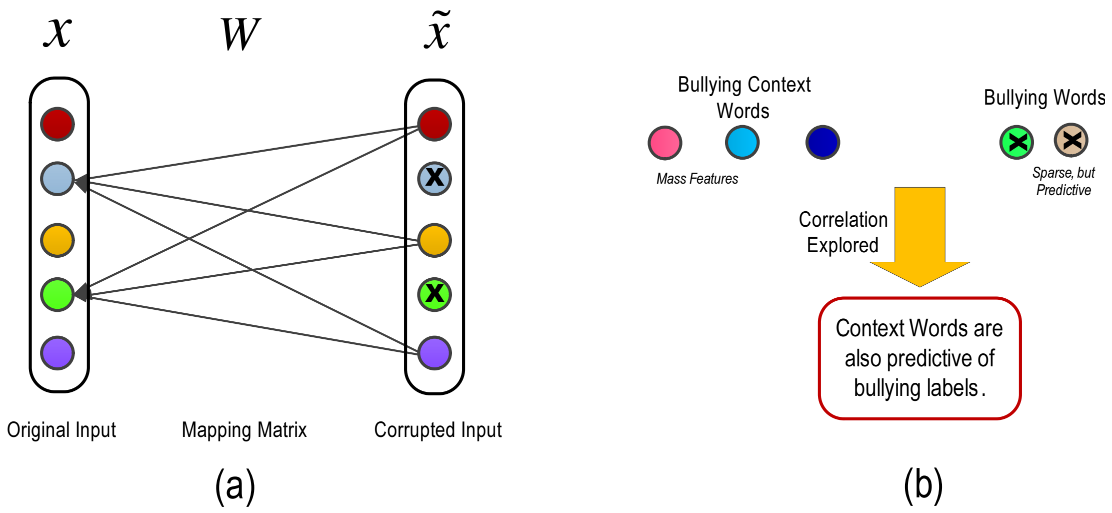


Fig. 1. Illustration of Motivations behind smSDA. In Figure 1(a), the cross symbol denotes that its corresponding feature is corrupted, i.e., turned off.

3.3 Construction of Bullying Feature Set

As analyzed above, the bullying features play an important role and should be chosen properly. In the following, the steps for constructing bullying feature set $\mathbb{Z}_b$ 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.

Layer One: 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. However, it is possible that expert knowledge is limited and does not reflect the current usage and style of cyberlanguage. Therefore, we expand the list of pre-defined insulting words, i.e. *insulting seeds*, based on word embeddings as follows:

Word embeddings use real-valued and low-dimensional vectors to represent semantics of words [32], [33]. The well-trained word embeddings lie in a vector space where similar words are placed close to each other. In addition, the cosine similarity between word embeddings is able to quantify the semantic similarity between words. Considering the Interent messages are our interested corpus, we utilize a well-trained word2vec model on a large-scale twitter corpus containing 400 million tweets [34]. A visualization of some word embeddings after dimensionality reduction (PCA) is shown in Figure 2. 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, since the word embeddings adopted here are trained in a large scale corpus from Twitter, the similarity captured by word embeddings can represent the specific language pattern. For example, the embedding of the misspelled word *fck* is close to the embedding of *fuck* so that the word *fck* can be automatically extracted based on word embeddings.

We extend the pre-defined *insulting seeds* based on word embeddings. For each insulting seed, similar words are extracted if their cosine similarities with insult seed exceed a predefined threshold. For bigram $w_l w_r$, we simply use an additive model to derive the corresponding embedding as follows:

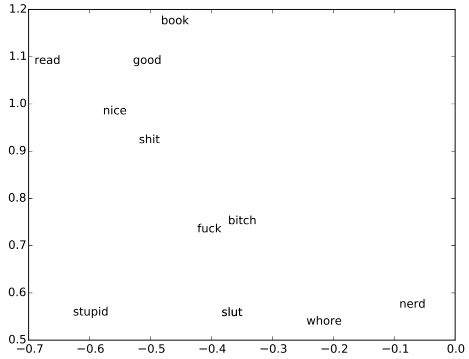


Fig. 2. Two dimensional visualization of our used word embeddings via PCA. Displayed terms include both bullying ones and normal ones. It shows that similar words are nearby vectors.

$$\mathbf{v}(w_l w_r) = \mathbf{v}(w_l) + \mathbf{v}(w_r) \tag{21}$$

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. The length of $\mathbb{Z}_b$ is k.

Subsequent Layers: we perform feature selection using Fisher score to select ''bullying'' features. Fisher score is an univariate metric reflecting the discriminative power of a feature [35], [36]. For the r^{th} feature, the corresponding Fisher score can be computed based on training data with labels:

$$F_r = \frac{\sum_{i=1}^{c} n_i(\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i {\sigma_i}^2} \tag{22}$$

where c denotes the number of classes and n_i represent the number of data in class i. μ and μ_i denote the mean of entire data and class i for the r^{th} feature, and σ_i is the variance of class i on r^{th} feature. After Fisher scores are estimated, features with top k scores are selected as ''bullying'' features, where ''bullying'' is generalized as discriminative.

3.4 smSDA for Cyberbullying Detection

In section 3.3, we propose the Semantic-enhanced Marginalized Stacked Denoising Auto-encoder (smSDA). In this subsection, we describe how to leverage it for cyberbullying detection. smSDA provides robust and discriminative representations The learned numerical representations can then be fed into Support Vector Machine (SVM). In the new space, due to the captured feature correlation and semantic information, the SVM, even trained in a small size of training corpus, is able to achieve a good performance on testing documents(this will be verified in the following experiments). The detailed steps of our model are provided below:

Assuming the first n_l posts are labeled and the corresponding vector of binary labels is $\mathbf{y} = \{y_1, \ldots, y_{n_l}\}$. The binary label 1 or 0 indicates the post is or is not a cyberbullying one. Here, $n_l \ll n$, which means the labeled posts have a small size. The bullying feature set $\mathbb{Z}_b$ is constructed in a layer-wise way. Based on prior knowledge, we construct a pre-defined bullying wordlist and compare it with the original vocabulary of the whole corpus $\mathbf{X}$. The words appearing in both the vocabulary and the bullying wordlist are selected as *insulting seeds*. The insulting seeds are then expanded and refined automatically via word embeddings, which defines the bullying features $\mathbb{Z}_b$ for layer one. The experiments in Section 4 will show that the construction of the set $\mathbb{Z}_b$ is very simple and efficient with litter human labor. For the subsequent layers, after obtaining the output of each layer, the set $\mathbb{Z}_b$ is updated using feature ranking with Fish score according to Eq. (22).

Based on predefined dropout probabilities for bullying features and other normal features p_b and p_n and the bullying feature set $\mathbb{Z}_b$, we compute these two expected matrices $E[\mathbf{P}]$ and $E[\mathbf{Q}]$ according to Eqs. (12) and (11), if the semantic dropout noise is adopted. When it comes to the unbiased semantic dropout noise, Eqs. (14) and (15) instead of Eqs. (12) and (11) are used to compute these two expected matrices. Then, we iteratively perform Eq. (21) for T_{max} times, where the initial value for $\mathbf{W}$ is calculated based on Eq. (20). When the mapping matrix is learned, the output of each layer is given according to Eq. (8). Due to the stacking structure, the output of L layers and the initial input are concatenated together to form the final representation $\mathbf{Z} \in \mathbb{R}^{d(L+1)\times n}$ following Eq. (9). It is clear that the new space has a dimension of $(L+1)d$. A linear SVM [37] is trained on the training corpus, i.e. the first n_l columns in $\mathbf{Z}$ and tested on the rest data samples.

3.5 Merits of smSDA

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

1) 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 autoencoder (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.
2) 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.
3) The sparsity constraint is injected into the solution of mapping matrix $\mathbf{W}$ 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 $\mathbf{W}$ 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.
4) 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.

4 Experiments

In this section, we evaluate our proposed semantic-enhanced marginalized stacked denoising auto-encoder (smSDA) with two public real-world cyberbullying corpora. We start by describing the adopted corpora and experimental setup. Experimental results are then compared with other baseline methods to test the performance of our approach. At last, we provide a detailed analysis to explain the good performance of our method.

4.1 Descriptions of Datasets

Two datasets are used here. One is from *Twitter* and another is from *MySpace* groups. The details of these two datasets are described below:

***Twitter* Dataset**: *Twitter* is ''a real-time information network that connects you to the latest stories, ideas, opinions and news about what you find interesting'' (https://about.twitter.com/). Registered users can read and post tweets, which are defined as the messages posted on *Twitter* with a maximum length of 140 characters.

The *Twitter* dataset is composed of tweets crawled by the public *Twitter* stream API through two steps. In Step 1, keywords starting with "bull" including "bully", "bullied" and "bullying" are used as queries in Twitter to preselect some tweets that potentially contain bullying contents. Retweets are removed by excluding tweets containing the acronym ''RT''. In Step 2, the selected tweets are manually labeled as bullying trace or non-bullying trace based on the contents of the tweets. 7321 tweets are randomly sampled from the whole tweets collections from August 6, 2011 to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2531682, IEEE Transactions on Affective Computing

August 31, 2011 and manually labeled[2]. It should be pointed out here that labeling is based on bullying traces. A bullying trace is defined as the response of participants to their bullying experience. Bullying traces include not only messages 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 research [38]. Some examples of bullying traces are shown in Figure 3. To preprocess these tweets, a tokenizer is applied without any stemming or stopword removal operations. In addition, some special characters including user mentions, URLS and so on are replaced by predefined characters, respectively. The features are composed of unigrams and bigrams that should appear at least twice and the details of preprocessing can be found in [8]. The statistics of this dataset can be found in Table 1.

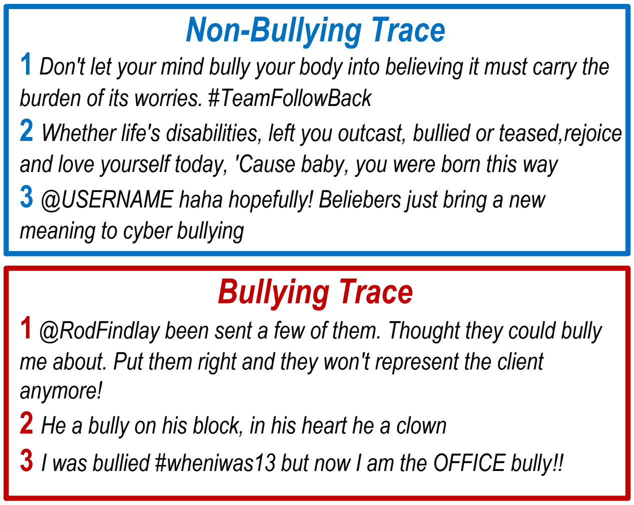
***MySpace* Dataset**: *MySpace* is another web2.0 social networking website. The registered accounts are allowed to view pictures, read chat and check other peoples' profile information.

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, each data sample is defined as a window of 10 consecutive posts and the windows are moved one post by one post so that we got multiple windows [39]. Then, three people labeled the data for the existence of bullying content independently. To be objective, an instance is labeled as cyberbullying only if at least 2 out of 3 coders identify bullying content in the windows of posts. 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, which are put into 11 packets. Some posts in *MySpace* are shown in Figure 4. Here, we focus on content-based mining, and hence, we only extract and preprocess the posts' text. The preprocessing steps of the *MySpace* raw text include tokenization, deletion of punctuation and special characters. The unigrams and bigrams features are adopted here. The threshold for negligible low-frequency terms is set to 20, considering one post occurred in a long conversation will occur in at least ten windows. The details of this dataset is shown in Table 1.

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. For *Twitter* dataset, we randomly select 800 instances, which accounts for 12% of the whole corpus, as the training data and the rest data samples are used as testing data. 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 testing. The process is repeated ten times to generate ten *sub*-datasets constructed from *MySpace* data. Finally, we have twenty *sub*-datasets, in which ten datasets are from *Twitter* corpus and another ten datasets are from *MySpace* corpus.

TABLE 1
Statistical Properties of the two datasets.

Statistics	*Twitter*	*MySpace*
Feature No.	4413	4240
Sample No.	7321	1539
Bullying Instances	2102	398


Non-Bullying Trace

1 Don't let your mind bully your body into believing it must carry the burden of its worries. #TeamFollowBack

2 Whether life's disabilities, left you outcast, bullied or teased,rejoice and love yourself today, 'Cause baby, you were born this way

3 @USERNAME haha hopefully! Beliebers just bring a new meaning to cyber bullying

Bullying Trace

1 @RodFindlay been sent a few of them. Thought they could bully me about. Put them right and they won't represent the client anymore!

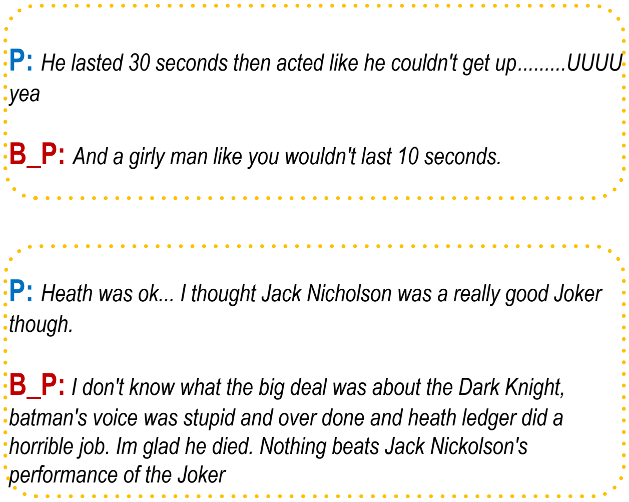
2 He a bully on his block, in his heart he a clown

3 I was bullied #wheniwas13 but now I am the OFFICE bully!!

Fig. 3. Some Examples from *Twitter* Datasets. Three of them are non-bullying traces. And the other three are bullying traces.

4.2 Experimental Setup

Here, we experimentally evaluate our smSDA on two cyberbullying detection corpora. The following methods will be compared.


P: He lasted 30 seconds then acted like he couldn't get up.........UUUU yea

B_P: And a girly man like you wouldn't last 10 seconds.

P: Heath was ok... I thought Jack Nicholson was a really good Joker though.

B_P: I don't know what the big deal was about the Dark Knight, batman's voice was stupid and over done and heath ledger did a horrible job. Im glad he died. Nothing beats Jack Nickolson's performance of the Joker

Fig. 4. Some Examples from *MySpace* Datasets. Two Conversions are Displayed and each one includes a normal post (P) and a bullying post (B_P) .

[2] The dataset: **bullyingV3.0**, has been kindly provided at http://research.cs.wisc.edu/bullying/data.html

[3] The dataset: **MySpace Group**, has been kindly provided at http://www.chatcoder.com/DataDownload

- BWM: Bullying word matching. If the message contains at least one of our defined bullying words, it will be classified as bullying.
- BoW Model: the raw BoW features are directly fed into the classifier.
- Semantic-enhanced BoW Model: This approach is referred in [12]. Following the original setting, we scale the bullying features by a factor of 2.
- LSA: Latent Semantic Analysis [18].
- LDA: Latent Dirchilet Allocation [20]. Our implementation of LDA is based on *Gensim*[4].
- mSDA: marginalized stacked denoising autoencoder [17].
- smSDA and smSDA$_\mathrm{u}$: semantic-enhanced marginalized denoising autoencoder that utilizes semantic dropout noise and unbiased one, respectively.

For LSA and LDA, the number of latent topics are both set to 100. In LDA, we set hyperparameter α for document topic multinomial and hyperparameter η for word topic multinomial to 1 and 0.01, respectively. For mSDA[5], the noise intensity is set to 0.5 and the number of layers for *Tweets* and *MySpace* datasets are both set to 2. Here, the number of layers is only set to be a moderate number instead of a large one, considering a large final dimension will impose a computational burden on the subsequent classifier training.

For our proposed methods including smSDA and smSDA$_\mathrm{u}$: the noise intensity and the number of layers are set to the same values as in mSDA to give a fair comparison. The bullying noise intensity is set to 0.8, which is larger than 0.5. The hyperparameters λ that controls the sparsity of the transformation matrix are set to 1 for all layers. The number of iteration step for solving lasso problems is set to 20. To construct the bullying features $\mathbb{Z}_b$ for the first layer, the negative word list containing 350 words is crawled[6], whose word cloud visualization is shown in Figure 5. The intersections between BoW features of our own corpus and the predefined bullying word list are firstly obtained. Then, as described in 3.3, they are extended and refined based on word embeddings to form the final bullying features. The threshold for cosine similarity is set to 0.8. The word cloud visualizations for the final bullying features in *Twitter* and *MySpace* datasets are shown in Figures 6 and 7, respectively. The bullying features used in Semantic-enhanced BoW Model are the same as those in smSDA.

Linear SVM [37] is then applied to the new feature space generated by the above mentioned approaches. In linear SVM, we search the best regularization parameter C from $\{0.0001, 0.001, 0.01, 0.1, 1\}$. To evaluate the performance of these methods on binary classification, classification accuracy is employed. Considering both datasets have the class imbalance problem, we also introduce F1-Score, which is a balance between precision and recall, to evaluate the performance of all compared approaches.

jerk
sand
coon
shiznit
pissed
scrote
cooter
sneeze
pecker
gaywad
gooch
cracker
shitdick
fellatio
fuckface
fuckboy
guido
buttfucka
muffdiver
cocksmoke
clitface
twatlips
peckerhead
dickfucker
shitter
dickhole
skeet
cock
dicksucker
shitcanned
motherfucker
cockmuncher
clusterfuck
dicksucking
carpetmuncher
unclefucker
cockmongruel
cumdumpster
twatwaffle
dickhead
cockjockey
cocksmoke
cockmaster
goddamnit
fuckhole
fucker
cumguzzler
dumbshit
shithole
nigger
wetback
wankjob
shitty
bitchass
fuckwit
assfuck
punta
lezzie
bullshit
panooch
fagtard
renob
boner
fucks
camel
piss
prick
nigga
jiz
gayfuckist
porchmonkey
jackersucker
blow
fudgepacker
testicle
poontang
fucknut
shitcunt
dickwad
shit
cuntrag
douche
gaytard
dumbass
faggot
gayass
honkey
pirate
dick
tard
bollox
asspirate
toe

Fig. 5. Word Cloud Visualization of the List of Words with Negative Affective.

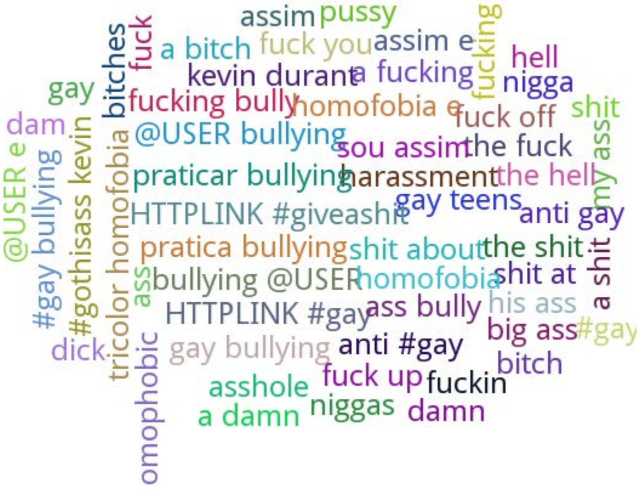


Fig. 6. Word Cloud Visualization of the Bullying Features in *Twitter* Datasets.

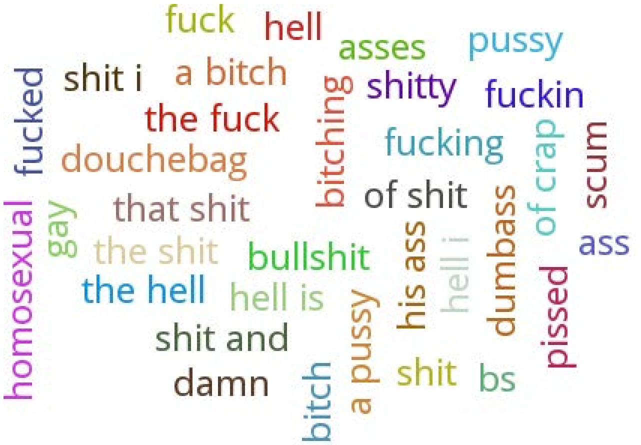


Fig. 7. Word Cloud Visualization of the Bullying Features in *MySpace* Datasets.

[4] https://radimrehurek.com/gensim/index.html

[5] The code has been kindly provided at http://research.cs.wisc.edu/bullying/data.html

[6] A collection of insulting words can be found in the website: http://www.noswearing.com/dictionary

4.3 Experimental Results

In this section, we show a comparison of our proposed smSDA method with six benchmark approaches on *Twitter* and *MySpace* datasets. The average results, for these two datasets, on classification accuracy and F1 score are shown in Table 2. Figures 8 and 9 show the results of seven compared approaches on all *sub*-datasets constructed from *Twitter* and *MySpace* datasets, respectively. Since BWM does not require training documents, its results over the whole corpus are reported in Table 2. It is clear that our approaches outperform the other approaches in these two *Twitter* and *MySpace* corpora.

The first observation is that semantic BoW model (sBow) performs slightly better than BoW. Based on BoW, sBoW just arbitrarily scale the bullying features by a factor of 2. This means that semantic information can boost the performance of cyberbullying detection. For a fair comparison, the bullying features used in our method and sBoW are unified to be the same. Our approaches, especially smSDA, gains a significant performance improvement compared to sBoW. This is because bullying features only account for a small portion of 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 correlation between normal features and bullying features by reconstructing corrupted data so as to yield robust features. In addition, Bullying Word Matching (BWM), as a simple and intuitive method of using semantic information, gives the worst performance. In BWM, the existence of bullying words are defined as rules for classification. It shows that only an elaborated utilization of such bullying words instead of a simple one can help cyberbullying detection.

We also compare our methods with two stat-of-arts text representation learning methods LSA and LDA. These two methods do not produce good performance on all datasets. This may be because that both methods belong to dimensionality reduction techniques, which are performed on the document-word occurrence matrix. Although the two methods try to minimize the reconstruction error as our approach does, the optimization in LSA and LDA is conducted after dimensionality reduction. The reduced dimension is a key parameter to determine the quality of learned feature space. Here, we fix the dimension of latent space to 100. Therefore, a deliberate searching for this parameter which may improve the performances of LSA and LDA and the selection of hyperparameter itself is another tough research topic. Another reason may be that the data samples are small (less than 2000) and the length of each Internet message is short (For *Twitter*, maximum length is 140 characters),and thus the constructed document-word occurrence matrix may not represent the true co-occurrence of terms.

Deep learning methods including mSDA and smSDA generally outperform other standard approaches. This trend is particularly prominent in F1 measure because cyberbullying detection problems are class-imbalance. The larger improvements on F1 score verify the performance of our approach further. Deep learning models have achieved remarkable performance in various scenarios with its own robust feature learning ability [22]. mSDA is able to capture the correlation between input features and combine the correlated features by reconstructing masking feature values from uncorrupted feature values. Further, the stacking structure and the nonlinearity contribute to mSDA's ability for discovering complex factors behind data. Based on mSDA, our proposed smSDA utilizes semantic dropout noise and sparsity constraints on mapping matrix, in which the efficiency of training can be kept. This extension leads to a stable performance improvement on cyberbullying detection and the detailed analysis has been provided in the following section.

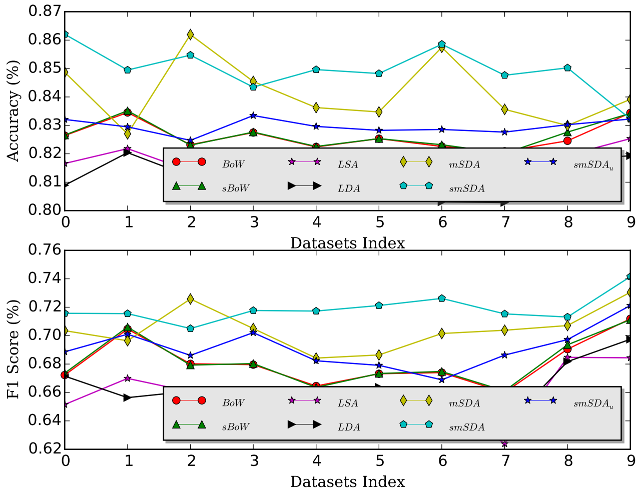


Fig. 8. Classification Accuracies and F1 Scores of All Compared Methods on *Twitter* Datasets.

We compare the performances of smSDA and smSDA$_{\mathrm{u}}$, which adopt biased semantic dropout noise and unbiased semantic dropout noise, respectively. The results have shown that smSDA$_{\mathrm{u}}$ performs slightly worse than smSDA. This may be explained by the fact that the unbiased semantic dropout noise cancels the enhancement of bullying features. As shown in Eq. (14), the off-diagonal elements in the matrix $\mathbf{x}_i\tilde{\mathbf{x}}_i^{\mathrm{T}}$ that are used to compute mapping weights are the same, which can not contribute to the reinforcement of bullying features.

4.4 Analysis of Semantic Extension

As shown in the section 4.3, the semantic extension can boost the performance on classification results for cyberbullying detection. In this section, we discuss the advantages of this extension qualitatively. In our proposed smSDA, because of the semantic dropout noise and sparsity constraints, the learned representation is able to discover the correlation between words containing latent bullying semantics. Table 3 shows the reconstruction terms of three example bullying words for mSDA and smSDA, respectively. In this example, one-hot vector is used as input, which represents a document containing one bullying word. Table 3 lists the reconstructed terms in decreasing order of their feature values, which represents the strength of their correlations with the input word. The results are obtained using one layer architecture without non-linear activation considering the raw terms directly correspond to

TABLE 2
Accuracies (%), and F1 Scores (%) for Compared Methods on *Twitter* and *MySpace* Datasets. The Mean Values are Given, respectively. Bold Face Indicates Best Performance.

Dataset	Measures	BWM	BoW	sBow	LSA	LDA	mSDA	smSDA$_u$	smSDA
Twitter	Accuracies	69.3	82.6	82.7	81.6	81.1	84.1	82.9	**84.9**
	F1 Scores	16.1	68.1	68.3	65.8	66.1	70.4	69.3	**71.9**
MySpace	Accuracies	34.2	80.1	80.1	77.7	77.8	87.8	88.0	**89.7**
	F1 Scores	36.4	41.2	42.5	45.0	43.1	76.1	76.0	**77.6**

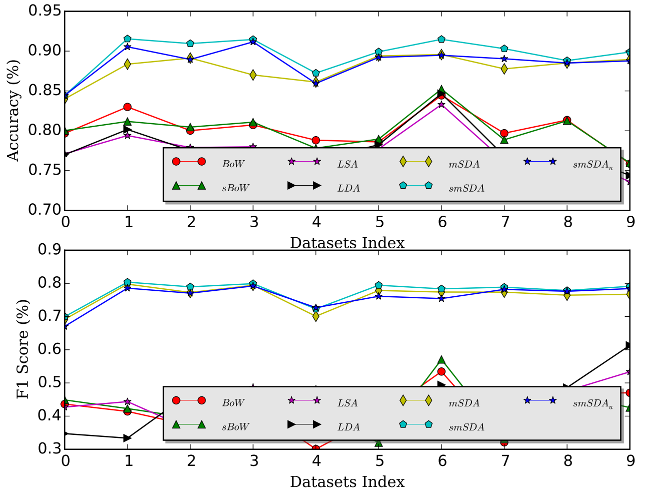


Fig. 9. Classification Accuracies and F1 Scores of All Compared Methods on *MySpace* Datasets.

TABLE 3
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).

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*

each output dimension under such a setting. It is shown that these reconstructed words discovered by smSDA are more correlated to bullying words than those by mSDA. For example, *fucking* is reconstructed by *because*, *friend*, *off*, *gets* in mSDA. Except *off*, the other three words seem to be unreasonable. However, in smSDA, *fucking* is reconstructed by *off*, *pissed*, *shit* and *of*. The occurrence of the term *of* may be due to the frequent misspelling in Internet writing. It is obvious that the correlation discovered by smSDA is more meaningful. This indicates that smSDA can learn the words' correlations which may be the signs of bullying semantics, and therefore the learned robust features boost the performance on cyberbullying detection.

5 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 learned representation by considering word order in messages.

Acknowledgments

We thank Junming Xu and Prof. Jerry Zhu for their *Twitter* datasets.

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 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.

[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 Conjugate Gradient-Related Methods*, pp. 92–100, 1996.

[30] 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, 2001.

[31] C. Vogel, *Computational Methods for Inverse Problems*. Society for Industrial and Applied Mathematics, 2002. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/1.9780898717570

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

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[34] F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle, "Named entity recognition for twitter microposts using distributed word representations," in *Proceedings of the Workshop on Noisy User-generated Text*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 146–153. [Online]. Available: http://www.aclweb.org/anthology/W15-4322

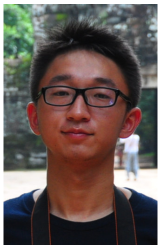
[35] T. H. Dat and C. Guan, "Feature selection based on fisher ratio and mutual information analyses for robust brain computer interface," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1. IEEE, 2007, pp. I–337.

[36] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[37] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[38] J. Sui, "Understanding and fighting bullying with machine learning," Ph.D. dissertation, THE UNIVERSITY OF WISCONSIN-MADISON, 2015.

[39] J. Bayzick, A. Kontostathis, and L. Edwards, "Detecting the presence of cyberbullying using computer software," in *Proceedings of the ACM WebSci'11*. Koblenz, Germany: ACM, June 2011, pp. 1–2.



Rui Zhao received the BEng in Measurement and Control from Southeast University, Nanjing, China, in 2012. He is currently pursuing the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

His current research interests include text mining and machine learning.

Kezhi Mao received his BEng, MEng and PhD from Jinan University, Northeastern University and Sheffield University in 1989, 1992 and 1998, respectively. He worked as a Lecturer at Northeastern University from March 1992 to May 1995, a Research Associate at University of Sheffield from April 1998 to September 1998, a Research Fellow at Nanyang Technological University from September 1998 to May 2001, an Assistant Professor at School of Electrical and Electronic Engineering, Nanyang Technological University from June 2001 to Sept 2005. He has been an Associate Professor since October 2005.

His areas of interests include computational intelligence, pattern recognition, text mining and knowledge extraction, cognitive science, and big data and text analytics.