RESEARCH ARTICLE OPEN ACCESS

Cberbullying Dection Using Semantic Enhanced Marginalized Denoising

Mr.B.S.Venkata Reddy[1], Mr.V.TataRao [2]
Department of Computer Science and Engineering
Raghu Engineering College(Autonomous), Andhra Pradesh,India

ABSTRACT

As a side result of increasingly popular social media, cyberbullying has emerged as a serious problem afflicting children, teens and immature adults. Machine learning techniques makes automatic detection of singling out messages in social media possible, and this could help to build a hale and hearty and safe social media upbringing. In this meaningful research area, one critical issue is robust and discriminative numerical representation learning of text messages. In this report, we suggest a new illustration, 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 morphological extension consists of semantic dropout noise and sparsity constraint, where the semantic dropout noise is projected based on domain knowledge and the word embedding technique. Our suggested method is able to exploit the hidden feature structure of bullying information and take a robust and discriminative representation of textual matter. Comprehensive experiments on two public cyber bullying corpora (Twitter and MySpace) are taken, and the results prove that our proposed approaches outperform other baseline text representation learning methods.

Keywords:- Cyberbullying Detection, Text Mining, Representation Learning, Stacked Denoising Autoencoders, Word Embedding

I. INTRODUCTION

SOCIAL Media, as fixed in [1], is ‘’a group of Internet-based applications that establish on the ideological and technical 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 along. However, social me-dia may have some side personal property such as cyberbullying, which may have pessimistic impacts on the life of people, especially children and teenagers.

Cyberbullying can be defined as antagonistic, intentional actions deliver the goods 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 takes place at school during face- to-face communication, cyberbullying on social media can take place anywhere at whatever time. For despotize, they are free to hurt their peers’ way of thinking because they do not need to face someone and can hide behind the information superhighway. For victims, they are easily exposed to torment for all of us, specially youth, are constantly connected to the Internet or social media. As described in [2], cyberbullying victimization rate ranges from 10% to 40%. In the United States, about 43% of teenagers were ever bullied on social media [3]. The same as traditional bullying, cyberbullying has negative, insidious and across-the-board impacts on children [4], [5], [6]. The results for victims under cyberbullying may even be tragic such as the occurrence of self-injurious behavior or suicides. One room to address the cyberbullying problem is to automatically notice and promptly report bullying messages so that proper steps can be adopted to prevent possible tragedies. Previous works on competition studies of bullying have shown that natural language processing and machine learning are powerful tools to study bullying [7], [8]. 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 demonstration learning for text communication. In fact, representation, learning of text is extensively read in text mining, information retrieval and natural speech processing (NLP). Suitcase-of-words (BoW) model is one commonly used model that each dimension corresponds to a condition. 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 depiction can be further processed for numerous language processing tasks. In cyberbullying detection, the statistical representation for Internet messages should be robust and particular. Even worse, the lack of sufficient high-quality training data, i.e., data sparsity make the issue more challenging.

Firstly, labeling data is labor intensive and time consuming. Secondly, cyberbullying is hard to identify and judge from a third view due to its intrinsic ambiguities. Thirdly, due to protection of Internet users and space to yourself issues, only a minor contribute to of messages are left on the Internet, and most bullying posts are edited. As a outcome, the trained classifier may not generalize well on testing messages that contain nonactivated but perceptive features.

Some plan of attacks have been suggested to take on these problems by integrating expert knowledge into feature learning. In addition, common sense knowledge was also used. 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]. Only a major restriction of these attacks is that the learned feature space still relies on the BoW assumption and may not be robust. In improver, the execution of these approaches relies on the quality of handcrafted features, which require extensive field knowledge.

In this paper, we investigate one deep learning method named stacked demonizing autoencoder (SDA) [15]. SDA stacks several dancing autoencoders and concatenates the output of each stratum as the learned representation. Each denoising autoencoder in SDA is trained to recover the input data from a debased version of it. This denoising process helps the autoencoders to learn a robust agency. In summation, each autoencoder layer is intended to take an increasingly abstract representation of the input [16]. We utilize semantic information to expand mSDA and develop Semantic-enhanced Marginalized Stacked Denoising Au- toencoders (smSDA). The semantic information consists of bullying words. An automatic extraction of bullying words based on word embeddings is proposed so that the involved human labor can be reduced.

During training of smSDA, we try to reconstruct bullying features from other normal words by identifying the latent structure, i.e. correlation, between bullying and normal speech. The intuition behind this thought 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 instance, in that location is a substantial correlation between bullying word fuck and normal word off since they frequently come together. If bullying messages do not take such obvious bullying features, such as fuck is often misspelled as folk, the correlation may help to reconstruct the bullying features from normal ones so that the bullying message can be found. It should be observed that introducing dropout noise has the effects of expanding the size of the dataset, including training data size, which helps alleviate the data sparsity problem. In addition, L1 formularization of the projection matrix is added to the objective function of each art in corner layer in our model to enforce the sparstiy of projection matrix, and this in turn facilitate the discovery of the most relevant terms for reconstruct bullying terms. The main contributions of our work can be summarized as follows:

* Our proposed Semantic-enhanced Marginalized S- tacked Denoising Autoencoder is able to learn ro- bust features from BoW representation in an effi- cient and effective way.
* The new feature space can improve the performance of cyberbullying de- tection even with a small labeled training corpus.
* Semantic information is incorporated into the re- construction process via the designing of semantic dropout noises and imposing sparsity constraints on mapping matrix. In our framework, high-quality semantic information, i.e., bullying words, can be extracted automatically through word embed.
* 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 structured as follows. In Section 2, some re- later work is introducing. The proposed Semantic enhanced Marginalized Stacked Denoising Auto encoder for cyber- bullying concealment is presented in Section 3. In Section 4, experimental results on several collections of cyberbullying data are illustrated. Finally, concluding remarks are provid- ed in Section 5.

II. RELATED WORK

This study points to determine a robust and discriminative text representation for cyberbullying detection. Text representation and automatic cyber bullying detection are both connected to our study. In the chase, we briefly review the previous study in these two countries.

1.1 Text Representation Learning

In text mining, information retrieval and natural language processing, effective numerical representation of syntactical units is a key issue. 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 turned out to be efficient and efficacious, the theatrical performance is often very sparse.

To address this problem, LSA applies Singular Value Decom position (SVD) on the word-document matrix for BoW model to derive a low-rank approximation. Each new feature is a linear amalgamation of all original features to alleviate the slenderness problem. The basic idea behind topic models is that word choice in a document will be determined by the topic of the document problematical.

Standardized to the approaches aforementioned, our proposed approach takes the BoW representation as the input. Nevertheless, our glide slope has some distinct merits. Firstly, the multilayers and nono of our model can ensure a deep learning architecture for text demonstration, which has been proven to be effective for learning high-level features [22]. Second, the applied abandon noise can make the learned demonstration more robust.

1.2 Cyberbullying Detection

With the increasing popularity of social media in recent years, cyberbullying has emerged as a serious problem afflicting children and immature adults. Previous fields of cyber bullying focused on broad studies and its payecological effects on victims, and were mainly led 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 on attracted the involvement of researchers. Several research areas including topic detection and affective analysis are closely related to cyberbullying detect. In machine learning-based cyberbullying detection, there are two issues: 1) text representation, learning to translate each post/message into a numerical vector and 2) classifier training. 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 debut of the sentiment and Contex Although the incorporation of the knowledge base can achieve a presentation improvement, the manufacture of a complete and general one is labor uncontrollable. The motivation behind this work is quite alike to that of our model to enhance bullying features.
Nevertheless, the scaling operation in [12] is quite arbitrary.

The weights for each extracted pattern need to be estimated based on the annotated training corpus, and hence the execution may not be warranted if the training corpus has a limited size. Huang et.al also considered social network features to learn the features for cyberbullying detection [9]. Different from these attacks, our proposed model can learn robust features by reconstructing the original information 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 discriminating.

1 Semantic-Enhanced MarginalizeStacked 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 $x \in R^d$. Then, the whole corpus can be denoted as a matrix: $X = [x_1, \ldots, x_n] \in R^{d \times n}$, where n is the number of available posts.

1.3 Marginalized Stacked Denoising Auto-encoder

Chen et.al proposed a modified version of Stacked Denoisin Auto-encoder that uses a linear instead of a non- linear projection so as to get a closed-form solution [17]. The basic idea behind demonizing auto encore is to recon- struct the original input from a corrupted one $\tilde{x}_1, \ldots, \tilde{x}_n$ with the goal of obtaining robust demonstration.

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:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{2n} tr\left(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}}\right)^{\top}\left(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}}\right) \quad (1)$$

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:

$$\mathbf{W} = \mathbf{P}\mathbf{Q}^{-1} \quad (2)$$

where $P = X\tilde{X}^T$ and $Q = \tilde{X}\tilde{X}^T$. 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. can be formulated as:

$$W = E[P]\, E[Q]^{-1} \quad (3)$$

where $E[P] = \Sigma n E \Sigma x i \tilde{x} T \Sigma$ and $E[Q]$

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 $S = XX^T$, the expected matrices can be computed as:

$$E[Q]_{i,j} = (1 - p)S_{i,j}$$

Where I and j denotes the indices of features. It can be understood that it is very efficient to compute W by marginalizing dropout noise in denoising auto-encoder. Later on the mapping weights W are computed, a nonlinear squashing function, such as a hyperbolic tangent function, can be applied to de- rival the turnout of the marginalized denoising auto-encoder:

$$\mathbf{H} = \tanh(\mathbf{W}\mathbf{X}) \quad (4)$$

Stacking Structure: Chen et.al [17] also proposed to apply stacking structures on marginalized denoising au- toencoder, in which the output of the $(k \quad 1)^{th}$ layer is fed as the input into the k^{th} layer. If we define the output of the k^{th} mDA as H_k and the original input as H_0 respectively, the mapping between two consecutive layers is given as:

$$\mathbf{H}_k = \tanh(\mathbf{W}_k\mathbf{H}_{k-1}) \quad (5)$$

where 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 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 X is the concatenation of the uncorrupted original input and outputs of all layers as follows:

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

1.4 Semantic Enhancement for mSDA

The advantage of purchasing the original input in mesa can be explained by feature co-occurrence info. As shown in Figure 1. (a), a demonizing auto encoder is trained to reconstruct these removed features values from the rest unaffected ones. It is shown that the learned demonstration is robust and can be regarded as a high level concept feature since the correlation information.

14.1 Semantic Dropout Noise

The dropout noise adopted in mSDA is an uniform distribution, where each feature has the same probability.

In cyber bullying 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 possibility on bullying words make emphatic the correlation between bullying features and normal ones. This kind of withdraw noise can be denoted as semantic dropout noise, because semantic information is applied to design dropout structure.

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. 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 partaken in by other bullying words contained in the 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 bullying features and normal features better and hence, facilitates cyberbullying detection.

Referable to the introduced semantic dropout noise, the expected matrices: E [P] and E [Q] will be computed slightly differently from ex. (5) and (6). Presuming we accept an available bullying words list and the corresponding facial appearance set Zb, the semantic dropout noise can be reported as the following prospect density occupation (PDF):

$$\text{sPDF}=p(\tilde{x}_d = x_d) = 1 - p_n \quad \text{if } d \in/ Z_b \qquad (6)$$

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

$$E[Q]_{i,j} = (1 - p_n)S_{i,j}$$

$$\text{If } i = j \;\&\; i \in/ Z_b,$$

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 . (6), the corrupted data is biased, i.e., E [X] E $\tilde{X}$. Here, we modified Eq. (10) to achieve an unbiased noise as follows:

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

the corrupted data is unbiased now. These two marginalized matrices are reformulated as:

$$E[Q]_{i,j}^{unbiased} = \begin{cases} \frac{1}{1-p_n} S_{i,j} & \text{if } i = j \;\&\; i \notin Z_b, \\ \frac{1}{1-p_b} S_{i,j} & \text{if } i = j \;\&\; i \in Z_b, \\ S_{i,j} & \text{if } i \neq j. \end{cases} \qquad (7)$$

$$E[P]_{i,j}^{unbiased} = S_{i,j} \qquad (8)$$

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

In mSDA, the mapping matrix W is learned to reconstrucst 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 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.

$$\mathbf{W} = \text{argmin} \frac{1}{} tr \; (\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}})^{\top}(\mathbf{X} - \mathbf{W}\tilde{\mathbf{X}}) + \lambda \|\mathbf{W}\| \quad (9)$$

where λ is a regularization parameter that controls the sparsity of W. The larger the λ is, the sparser the mapping matrix W is. The solution to Eq. (9) is a very mature math problem: sparse least squares optimization. [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 approximat substituting this approximation Eq. (9) into the objective function Eq. (9), we yield an formulation similar to a Ridge Regression Problem [31], and the iteration steps to solve W is given as:

$$\mathbf{W}_k = \tilde{\mathbf{X}}^{\top}\mathbf{X} \; \tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}} + \lambda \, \text{diag}(|\mathbf{W}_{k-1}|)^{-1} \qquad (10)$$

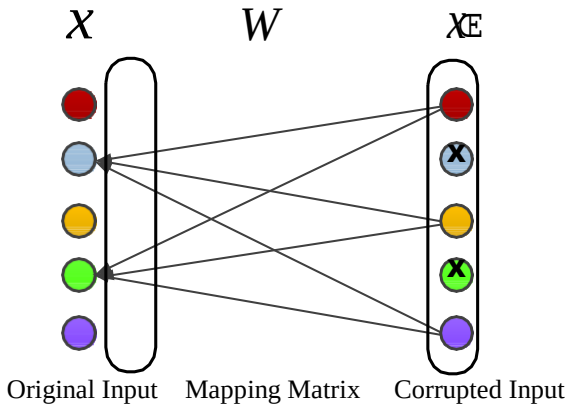
where diagram denotes the diagonal elements of a matrix, W_k and W_{k-1} denote the current step and the previous step estimations for mapping matrix W, respectively. It is clear that the Eq. (18) can be easily formulated when the noise distribution is marginalized. Similar to Eq. (4), Eq. (10) can be written as:

$$Wk = E[P]\Sigma E[Q] + \lambda \, \text{diag}(|Wk-1|)-1\Sigma-1 \quad (11)$$

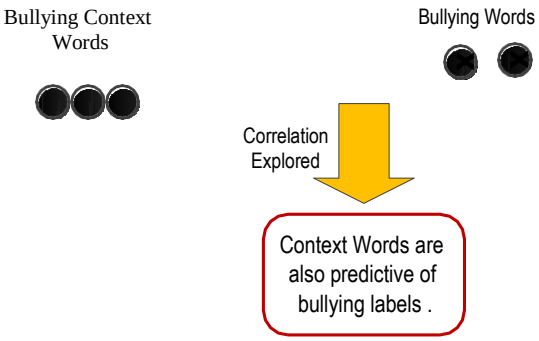
To speed up the convergence process, the initialization for W can be set to the L2 penalized solution for Eq. (2) as follows:

$$W0 = E[P]\Sigma E[Q] + \lambda I \, \Sigma-1 \quad (12)$$

where 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.



(a)



(b)

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.

1.5 Construction of Bullying Feature Set

As analyzed in a higher place, the bullying features play an important role and should be taken properly. In the going all out, the steps for throw together a bullying feature set Zb are given, in which the first layer and the other layers are addressed on an individual base. For the first layer, expert knowledge and word cascade are used. For the other layers, discerning feature selection is conducted.

Layer One: firstly, we create a list of words with negative affect, not including give your word words and filthy words. And so, we equate the word list with the BoW features of our own corpus, and involve the cloverleaf as bullying features.

However, it is possible that expert knowledge is limited and does not reflect the current usage and style of cyberlanguage. Word bombardings use real-valued and low-dimensional vectors to represent significs of words [32], [33]. In addition, the casing correspondence between word bombardings is able to quantify the semantic connection between words. Making an allowance for the Interent messages are our fascinated corpus, we utilize a well trained word2vec model on a large-scale twitter corpus containing 400 million tweets [34]. A mental picture of some word embeddings after dimensionality reduction (PCA) is shown in Figure 2. It is observed that curse words form diverse cluster, which are also far away from normal words. Even discourteous words are located at different region due to different word matter of course and insulting expressions. 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.

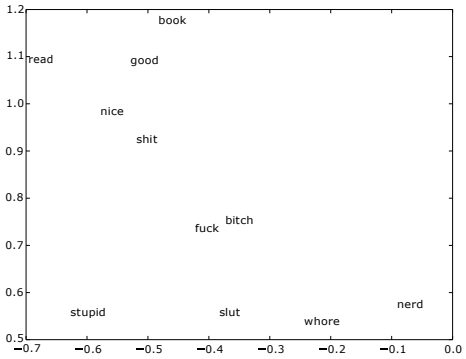


Fig. 2. Two dimensional visualization of our used word embeddings via PCA. Displayed terms include both bullying ones and normal ones. It shows that similar words are nearby vectors.

Tracted if their cosine synonymity with rudeness seed exceed a profaned threshold. For bigram $w_l w_r$, we simply use an chemical addition model to derive the corresponding embedding as follows:

$$v(wlwr) = v(wl) + v(wr) \qquad (13)$$

Finally, the constructed maltreatment features are used to train the first layer in our wished for smSDA. It includes two parts: one is the original wounding seeds based on domain acquaintance and the other is the unmitigated bullying words via word embeddings. The length of Zb is k.

Subsequent Layers: we perform feature mixture using Fisher scores to select ''bullying' facial appearance. For the rth feature, the analogous Fisher score can be computed based on training data with labels:

$$r_r = \frac{\sum_{c_i=1} n_i(\mu_i - \mu)^2}{\sum_{c_i=1} n^i \sigma^{i2}}$$

Where c hangs a sign on the quantity of classes and I be a symbol of the number of data in class I. After Fisher scores are projected, features with top k score are selected as ''bullying' features, where ''bullying' is wide-ranging as whole story.

1.6 smSDA for Cyberbullying Detection

In section 3.3, we propose the morphological enhanced marginalizing Stacked demonizing Auto encoder (smSDA). In this subsection, we describe how to leverage it for cyberbullying recognition. In the tender place, due to the captured feature relationship and semantic in sequence, the SVM, even experienced in a small size of working out corpus, is able to achieve a full reading on testing documents.

Taking over the first nail posts are labeled and the corresponding vector of binary labels is why = $y1, ., \ldots, y_n$**1** . The binary label 1 or 0 indicates the post is or is not a cyberbullying one. Here, no, no, which means the labeled posts have a humble size. Based on prior knowledge, we construct a pre-defined bullying wordlist and compare it with the original vocabulary of the whole corpus X. The words punch the clock in both the vocabulary and the bullying wordlist are selected as insulting seeds. The insulting seeds are then extended and refined without human intrusion via word embedding, which defines the bullying features Zb for layer one. For the subsequent layers, after holding the production of each stratum, the set Zb is modernized using feature ranking with Fish score according to Eq. (22).

Based on predefined dropout probabilities for bullying distinctiveness and other normal features pub and on and the bullying aspect set Zb, we think these two expected matrices E [P] and E [Q] according to Eqs. (12) and (11), if the semantic dropout noise is adopted. When the square matrix is learned, the output of each layer is made concede to Eq. (8). Referable to the stacking structure, the output of L layers and the initial input are link together together to organize the final illustration Z Rd (L+1) ×n following Eq. (9).

1.7 Merits of smSDA

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

1) Most cyberbullying detection methods rely on the BoW model. Referable to the sparsity problems of both data and distinctiveness, the classify may not be smarten up very well. In SDA, the feature correlation is explored by the reconstruction of tarnished information. The learned stout feature illustration can then sponsor the training of clarifiers and finally improve the taxonomy accuracy. In sum, the altered form of data in SDA actually brings to pass reproduction data to make bigger data size, which alleviate the small size problem of education data.

2) For cyberbullying problem, we design, well-formed abandon noise to call inattention to bullying features in the new feature space, and the yielded new demonstration is thus more show favoritism for cyberbullying detection.

3) The sparsity constraint is injected into the solution of scaling matrix W for each layer, taking each word is only interconnected to a infinitesimal part of the whole terminology.

2 Experiments

In this section, we evaluate our wished for semantic enhanced marginalized stacked denoising auto-encoder (smSDA) with two public real world cyber bullying corpora. We begin by keying out the go down the line corpora and experiment tool setup.

2.1 Descriptions of Datasets

Two datasets are used here. Ace is from chirrup and another is from spacy groups. The details of these two datasets are described below:

Twitter Dataset: Twitter is ''a real-time information network that joins you to the latest tales, ideas, Opin- ions and news about what you discover 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. Re- tweets 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 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 mes- sages about direct bullying attack, but also messages about reporting a bullying experience, revealing self as a victim et. al. Thus, bullying traces far exceed the incidents of cyberbullying. Some examples of bullying traces are shown in Figure 3. To preprocess these tweets, a tokenizer is used without any stemming or stopword removal operations. The features are composed of unigrams and bigrams that should come out 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 seen as a conversation around one subject. verifiable to the partaken personality behind cyberbullying, each data sample is defined as a window of 10 repeated posts and the windows are moved one post by one post so that we got multiple windows [39]. To be objective, an illustration is labeled as cyberbullying only if at least 2 out of 3 coders identify bullying content in the windows of offices. The raw text for these data, as XML files, have been kindheartedly provided by Kontostathis et.al[3]. Here, we focus on content based mining, and hence, we just extract and preprocess the posts' text. The preprocessing steps of the MySpace raw text include tokenization, dualtone of punctuation and extra qualities. The unigrams and bigrams features are assumed here. The threshold for negligible low-frequency terms is set to 20, considering one post occurred in a long conversation will take place in at least ten windows. The details of this dataset is shown in Table 1. Since there were no stock splits of training vs. test datasets in our adopted Twitter and MySpace corpora, we require to define the breeding and testing data sets.As analyzed above that the lack of labeled instruction corpus hinders the development of involuntary 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 financial statement for 12% of the whole corpus, as the training data and the rest data samples are used as testing data.

TABLE 1
Statistical Properties of the two datasets.

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

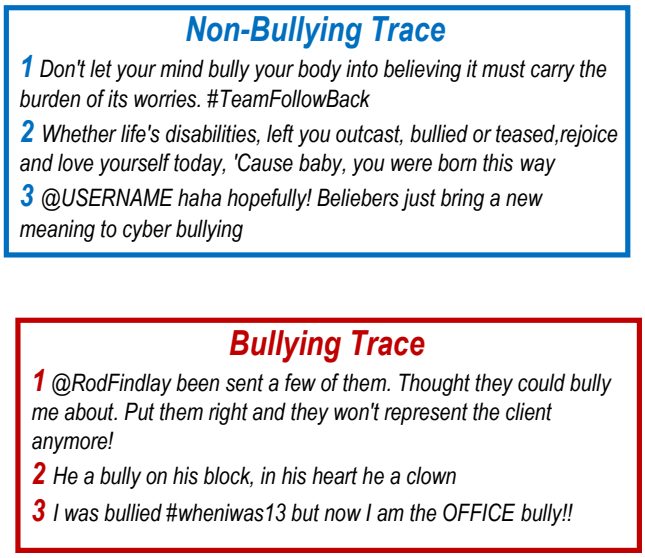


Fig. 3. Some Examples from Twitter Datasets. Three of them are non-bullying traces. And the other three are bullying traces.

The procedure is repeated ten times to generate ten sub-datasets construct from MySpace data. Lastly, we have twenty sub-datasets, in which ten data sheets are from Twitter corpus and another ten datasets are from MySpace corpus.

2.2 Experimental Setup

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

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) .

* BWM: Bullying word matching. If the message contains at least one of our amorphous bullying words, it will be reckon as bullying.
* BoW Model: the raw BoW features are directly fed into the classify.
* Semantic-enhanced BoW Model: This approach is cited in [12]. Sticking to the original mount, 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 autoen- coder [17].
* SmSDA and smSDAu: semantic-enhanced marginalized denoising autoencoder that utilizes semantic dropout noise and unbiased one, severally.

For LSA and LDA, the number of latent topics are both set to 100. 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. Hither, the number of layers is only set to be a moderate number instead of a big one, seeing a large, final dimension will impose a compotation burden on the subsequent classifier training.

For our suggested methods, including smSDA and smSDAu: the noise intensity and the number of layers are set to the same values as in mSDA to give a just comparison. The hyperparameters λ that controls the sparsity of the transformation matrix are set to 1 for all layers. The number of monotony step for solving lasso problems is set to 20.

To construct the bullying features Z_b for the first layer, the negative word list containing 350 words is crawled[6], whose word cloud revelation is shown in Figure 5. The cloverleaf between BoW features of our own corpus and the predefined bullying word list are firstly beg borrow or steal. Then, as described in 3.3, they are extended and refined based on word embeddings to form the final. Inward eye for the final bullying features in *Twitter* and *MySpace* datasets are shown in Figures 6 and 7,respective- ly.

Linear SVM [37] is then given to the new feature space generated by the above mentioned attacks. In linear SVM, we look for the best regularization parameter C from 0.0001, 0.001, 0.01, 0.1, 1. To evaluate the performance of these ways and means on binary classification, classification accuracy is employed. Considering both datasets have the class disproportion problem, we also introduce F1-Score, which is a balance between precision and recall, to evaluate the piece of all compared approaches.

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

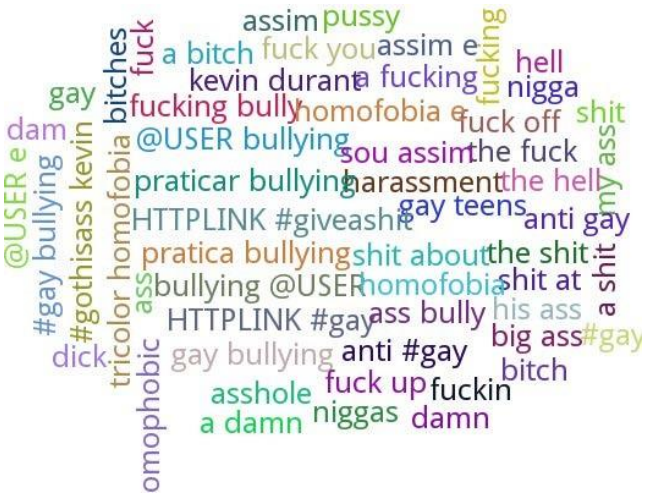


Fig. 6. Word Cloud Visualization of the Bullying Features in *Twitter* Datasets

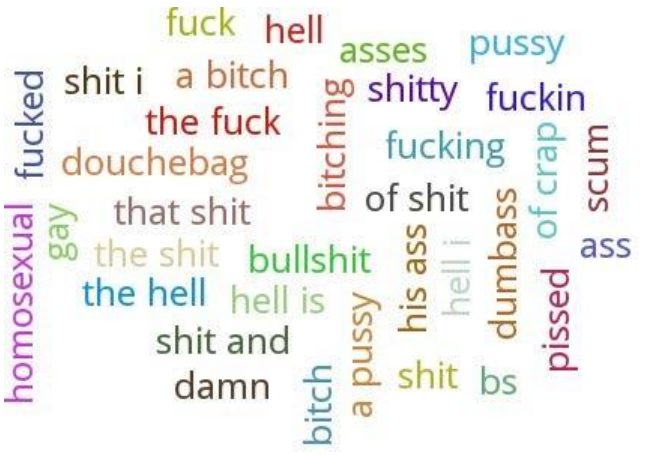


Fig. 7. Word Cloud Visualization of the Bullying Features in MySpace Datasets.

2.3 Experimental Results

In this segment, we present a comparison of our proposed smSDA method with six point of reference approaches on Twitter and MySpace datasets. Since BWM does not require training documents, its results over the whole corpus are reported in Table 2. It is clear that our approaches go one better than the other ingress in these two Twitter and MySpace corpora.

The first notice is that the semantic BoW model (show) performs slightly better than BoW. Based on BoW, sBoW just without rhyme or reason scale the bullying features by a factor of 2. This is because bullying features only account for a humble portion of all features used. It is hard to learn robust features for small training data by intensifying each bullying features extent. In addition, Bullying Word Matching (BWM), as a simple and intuitive method of using morphological information, gives the worst performance.

We also compare our methods with two statet-of-arts text representation, learning methods LSA and LDA. Although the two methods try to minimize the reconstruction error as our come within reach of does, the optimization in LSA and LDA is conducted after magnitude reduction. The reduced dimension is a key constraint to specify the character of learning 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 re- markable performance in various scenarios with its own robust feature learning ability [22]. mSDA is able to cap- ture the correlation between input features and combine the correlated features by reconstructing masking feature values from uncorrupted feature values. Further, the stack- ing structure and the nonlinearity contribute to mSDA's ability for discovering complex factors behind data.

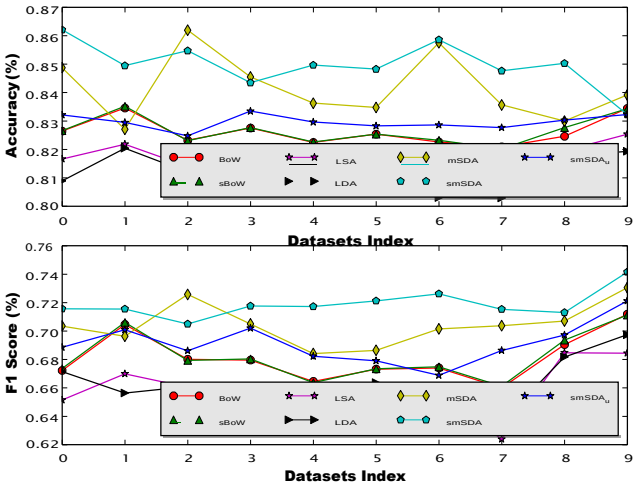


Fig. 8. Classification Accuracies and F1 Scores of All Compared Meth- ods on Twitter Datasets.

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 de- tection and the detailed analysis has been provided in the following section.

We compare the realization of smSDA and $smSDA_u$, which adopt biased semantic dropout noise and unbi- ased semantic dropout noise, respectively. This may be explicated by the fact that the dispassionate semantic dropout noise cancels the augmentation of bullying features. As shown in Eq. (14), the off-diagnosel elements in the matrix $x_i\tilde{x}^T$ that are used to compute mapping weights are the same, which can not contribute to the reinforcement of bullying features.

2.4 Analysis of Semantic Extension

As indicated in the section 4.3, the morphological conservatory can boost the performance on organization results for cyberbullying detection. In this section, we discuss the advan- tages 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, correspondingly. Table 3 lists the reconstructed terms in diminishing order of their feature values, which represents the strength of their correlation with the input word. The results are obtain using one layer construction without non-linear activation considering the raw terms directly keep up a correspondence to each output dimension under such a setting.

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**

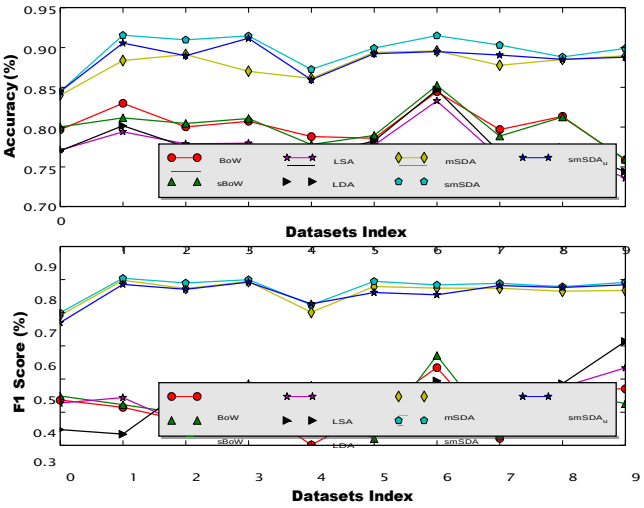


Fig. 9. Classification Accuracies and F1 Scores of All Compared Meth- ods on MySpace Datasets

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*

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).

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. 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.

3 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 spar- sity, we have developed semantic-enhanced marginalized denoising autoencoder as a specialized representation learn- ing 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.

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 psy- chosomatic 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 Compu- tational Linguistics, 2012, pp. 656–666.

[9] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in Proceedings of the 3rd Inter- national Workshop on Socially-Aware Multimedia. ACM, 2014, pp. 3–6.

[10] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7, 2009.

[11] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detec- tion of textual cyberbullying." in The Social Mobile Web, 2011.

[12] V. Nahar, X. Li, and C. Pang, "An effective approach for cy- berbullying detection," Communications in Information Science and Management Engineering, 2012.

[13] M. Dadvar, F. de Jong, R. Ordelman, and R. Trieschnigg, "Im- proved cyberbullying detection using gender information," in Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop (DIR2012). Ghent, Belgium: ACM, 2012.

[14] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Im- proving cyberbullying detection with user context," in Advances in Information Retrieval. Springer, 2013, pp. 693–696.

[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," The Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.

[16] P. Baldi, "Autoencoders, unsupervised learning, and deep archi- tectures," Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7, p. 43, 2012.

[17] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized de- noising autoencoders for domain adaptation," arXiv preprint arX- iv:1206.4683, 2012.

[18] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse processes, vol. 25, no. 2-3, pp. 259–284, 1998.

[19] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceed- ings 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 se- mantic 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 ground- s?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 chil- dren get bullied? a prospective cohort study on the relationship between bullying and health-related symptoms," Pediatrics, vol. 117, no. 5, pp. 1568–1574, 2006.

Mr. B. S. Venkata Reddy working as Assistant Professor in the Department of Computer Science and Engineering, Raghu Engineering College(Autonomous),Visakhapatnam, Andhrapradesh. His research interest is on Data Science, Big Data and Business Analytics, Cyber security, Social Media Marketing. He is pursuing part time Ph.D from AMET, Cheinnai. In connection with NGOs, Startups , Training Institutions, Colleges ,etc., he has certified various Technical Courses and Trained more than 10,000 students from 1998 to till date . He has membership in various organizations. He has a zeal to work for the development of the community with various innovative, creative and unique activities and Programs. His basic motto is to serve the nation through the people of the community.

Mr. V.TataRao working as Assistant Professor(Ratified Faculty by JNTUK) in the Department of Computer Science and Engineering, Raghu Engineering College (Autonomous), Visakhapatnam, Andhrapradesh. His research interest on Network security and Cyber security. He has obtained his B.Tech(CSE) from JNTUH, M.Tech(CSE) from JNTUK.