Cyberbullying Detection Based on Text Representation

K. Ruba Soundar¹, P. Ponesakki²
Head of the Department¹, PG scholar²
P.S.R Engineering College, Sivakasi, Tamil Nadu, India

Abstract:
As a side effect of increasingly popular social media, cyberbullying has emerged as a serious problem affecting 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 auto encoder. 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 cyber bullying 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 Auto encoders, Word Embedding

I. INTRODUCTION

Social Media, as defined as “a group of Internet based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation 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 victims, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or social media. As reported in cyberbullying victimization rate ranges from 10% to 40%. In the United States, approximately 43% of teenagers were ever bullied on social media. The same as traditional bullying, cyberbullying has negative, insidious and sweeping impacts on children. The outcomes for victims under cyberbullying may even be tragic such as the occurrence of self-injurious behavior or suicides. One way to address the cyberbullying problem is to automatically detect and promptly report bullying messages so that proper measures can be taken to prevent possible tragedies. Previous works on computational studies of bullying have shown that natural language processing and machine learning are powerful tools to study bullying. Cyberbullying detection can be formulated as a supervised learning problem. A classifier is first trained on a cyberbullying corpus labeled by humans and the learned classifier is then used to recognize a bullying message. Three kinds of information including text, user demography, and social network features are often used in cyberbullying detection. Since the text content is the most reliable, our work here focuses on text-based cyberbullying detection. In the text-based cyberbullying detection, the first and also critical step is the numerical representation learning for text messages. In fact, representation learning of text is extensively studied in text mining, information retrieval and natural language processing (NLP). Bag-of-words (BoW) model is one commonly used model that each dimension corresponds to a term. Latent Semantic Analysis (LSA) and topic models are another popular text representation models, which are both based on Bow models. By mapping text units into fixed-length vectors, the learned representation can be further processed for numerous language processing tasks. Therefore, the useful representation should discover the meaning behind text units. In cyberbullying detection, the numerical representation for Internet messages should be robust and discriminative. Since messages on social media are often very short and contain a lot of informal language and misspellings, robust representations for these messages are required to reduce their ambiguity. Even worse, the lack of sufficient high-quality training data, i.e., data sparsity make the issue more challenging. Firstly, labeling data is labor intensive and time consuming. Secondly, cyberbullying is hard to describe and judge from a third view due to its intrinsic ambiguities. Thirdly, due to protection of Internet users and privacy issues, only a small portion of messages are left on the Internet and most bullying posts are deleted. As a result, the trained classifier may not generalize well on testing messages that contain no activated but discriminative features. The goal of this present
study is to develop methods that can learn robust and
discriminative representations to tackle the above problems in
cyberbullying detection. Some approaches have been proposed to
tackle these problems by incorporating expert knowledge into
feature learning. Yin et al. proposed to combine BoW features,
sentiment features and contextual features to train a support
vector machine for online harassment detection [10]. Dinakaret.
al utilized label specific features to extend the general features,
where the label specific features are learned by Linear
Discriminative Analysis. In addition, common sense knowledge
was also applied. Nahar et al. presented a weighted TF-IDF
scheme via scaling bullying-like features by a factor of two.
Besides content-based information, Maral et al. proposed to
apply users’ information, such as gender and history messages,
and context information as extra features. But a major limitation
of these approaches is that the learned feature space still relies
on the BoW assumption and may not be robust. In addition, the
performances of these approaches rely on the quality of hand-
crafted features, which require extensive domain knowledge.
One deep learning method named stacked enoising autoencoder
(SDA). SBA 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 auto
encoders to learn robust representation. In addition, each auto
encoder layer is intended to learn an increasingly abstract
representation of the input. To develop new text representation
model based on a variant of SDA: marginalized stacked
denoising autoencoders (mSDA) which adopts linear instead of
nonlinear projection to accelerate training and marginalizes
infinite noise distribution in order to learn more robust
representations. We utilize semantic information to expand
mSDA and develop Semantic-enhanced Marginalized Stacked
Denoising Auto encoders (smSDA). The semantic information
consists of bullying words. An automatic extraction of bullying
words based on word 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 o 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 sparsity 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. The Semantic-enhanced
Marginalized Stacked Denoising Auto-encoder for cyberbullying
detection is presented in experimental results on several
collections of cyberbullying data are illustrated.

II. 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) and
topic models BoW model represents a document in a textual
corpus using a vector of real numbers indicating the occurrence
of words in the document. Although BoW model has proven to
be efficient and effective, the representation is often very sparse.
To address this problem, LSA applies Singular Value
Decomposition (SVD) on the word-document matrix for BoW
model to derive a low-rank approximation. Each new feature is a
linear combination of all original features to alleviate the
sparsity problem. Topic models, including Probabilistic Latent
Semantic Analysis and Learning Dirichlet Allocation are also
proposed. The basic idea behind topic a model 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. Second, the applied dropout noise
can make the learned representation more robust. On mapping
matrix in each layer and this will in turn produce more
discriminative representation, including bullying words and
sparsity constraint imposed on mapping matrix in each layer and
Cyberbullying Detection

With the increasing popularity of social media in recent years, cyberbullying has emerged as a serious problem affecting children and young adults. Previous studies of cyberbullying focused on extensive surveys and its psychological effects on victims, and were mainly conducted by social scientists and psychologists. Although these efforts facilitate our understanding for 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. As an introductory work; they did not develop specialized models for cyberbullying detection. Yin et al. proposed to combine BoW features, sentiment feature and contextual features to train a classifier for detecting possible harassing posts. The introduction of the sentiment and contextual features has been proven to be effective. Dimakar et al. used Linear Discriminative Analysis to learn label specific features and combine them with BoW features to train a classifier. The performance of label-specific features largely depends on the size of training corpus. In addition, they need to construct a bully space knowledge base to boost the performance of natural language processing methods. Although the incorporation of knowledge base can achieve a performance improvement, the construction of a complete and general one is labor-consuming. Nahar et al. proposed to scale bullying words by a factor of two in the original BoW features. The motivation behind this work is quite similar to that of our model to enhance bullying features. However, the scaling operation is quite arbitrary. Paszynski et al. searched sophisticated patterns in a brute-force way. The weights for each extracted pattern need to be calculated based on annotated training corpus, and thus the performance may not be guaranteed if the training corpus has a limited size. Besides content-based information, Maral et al. also employ users’ information, such as gender and history messages, and context information as extra features. Huang et al. also considered social network features to learn the features for cyberbullying detection. The shared deficiency among these aforementioned approaches is constructed text features are still from Bow representation, which has been criticized for its inherent over-sparsity and failure to capture semantic structure. Different from these approaches, our proposed model can learn robust features by reconstructing the original data from corrupted data and introduce semantic corruption noise and sparsity mapping matrix to explore the feature structure which are predictive of the existence of bullying so that the learned representation can be discriminative.

This will in turn produce more discriminative representation. 2.2

Marginalized 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 exist for EuR he’ Peeholen for us dusee as a matrix: \( X = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n} \), where \( n \) is the number of available posts. We next briefly review the marginalized stacked de-noising 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. The basic idea behind denoising auto-encoder is to re-construct the original input from a corrupted one \( I \), with the goal of obtaining robust representation.

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

\[
W = \underset{W}{\text{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \| x_i - W \hat{x}_i \|^2 \tag{1}
\]

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

\[
W = \underset{W}{\text{argmin}} \frac{1}{2n} \text{tr} \left[ (X - W \hat{X})^T (X - W \hat{X}) \right] \tag{2}
\]

where \( X = [x_1, \ldots, x_n] \) is the corrupted version of \( X \). It is easily shown that Eq. (2) is an ordinary least square problem having a closed-form solution.

\[
W = PQ^{-1} \tag{3}
\]

where \( P = XXT \) and \( Q = XXT \). In fact, this corruption an 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:

\[
W = \text{E}[P] \text{E}[Q]^{-1} \tag{4}
\]

Where \( \text{E}[P] = \sum_{i=1}^{n} \text{E}[x_i x_i^T] \) and \( \text{E}[Q] = \sum_{i=1}^{n} \text{E}[x_i x_i^T] \). These expected matrices can be computed based on noise distribution. In 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 \( S = XX^T \), the expected matrices can be computed as:

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

\[ E [P]_{i,j} = (1-P)S_{i,j} \] (6)

where \( i \) and \( j \) denotes the indices of features. It can be seen that it is very efficient to compute \( W \) by marginalizing dropout noise in denoising auto-encoder. After the mapping weights \( W \) 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:

\[ H = \tanh(WX) \] (7)

**Stacking Structure:** Chen et al 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 layer as \( H_k \) and the original input as \( X \) respectively, the mapping between two consecutive layers is given as:

\[ H_k = \tanh(W_kH_{k-1}) \] (8)

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 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:

\[ Z = [X_{H_1} \ldots X_{H_L}] \] (9)

### 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 four ways, 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 chiflk"l, this 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" is 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 bullying features and normal features better and hence, facilitates cyberbullying detection. Due to the introduced semantic dropout noise, the expected matrices \( E [P] \) and \( E [Q] \) will be computed slightly different from Eqs. (5) and (6). Assuming we have an available bullying words list and the corresponding features set \( Z_b \), the semantic dropout noise can be described as the following probability density function (PDF):

\[ \text{PDF}\text{=} \begin{cases} \text{p(xd}=0) & \text{- pn} \text{if d } \notin \text{ Zb,} \\ \text{p(xd} = \text{zd)} & \text{- Pb} \text{if d } \in \text{ Zb,} \end{cases} \] (10)

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

\[ E [Q]_{i,j} = \begin{cases} (1 - p_n)S_{i,j} & \text{if } i = j \& i \notin Z_b, \\ (1 - p_n)^2S_{i,j} & \text{if } i \neq j \& \{i, j\} \cap Z_b = \emptyset, \\ (1 - p_b)(1 - p_n)S_{i,j} & \text{if } \{i, j\} \notin Z_b \cap \{i, j\} \cap Z_b, \\ (1 - p_b)^2S_{i,j} & \text{if } i \neq j \& \{i, j\} \in Z_b, \\ (1 - p_b)S_{i,j} & \text{if } i = j \& i \in Z_b. \end{cases} \] (11)

And

\[ E [P]_{i,j} = \begin{cases} (1 - p_n)S_{i,j} & \text{if } j \cap Z_b = \emptyset, \\ (1 - p_b)S_{i,j} & \text{if } j \cap Z_b \neq \emptyset. \end{cases} \] (12)

where \( p_n \) and \( p_b \) are the probabilities of bullying features and normal features to be set to zero respectively, and \( p_b > p_n \). Here, \( p_n \) and \( p_b \) are both tunable hyper parameters for our proposed smSDA.

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

\[ \text{PDF}_\text{unbiased} = \begin{cases} \text{p(xd} = 0) & \text{= pn} \text{ if d } \notin \text{ Zb,} \\ \text{p(xd} = \text{zd}) & \text{= 1-pn} \text{ if d } \notin \text{ Zb,} \\ \text{p(xd} = \text{zd}) & \text{= pb} \text{ if d } \notin \text{ Zb,} \end{cases} \] (13)

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

\[ E [Q]_{i,j}^{\text{unbiased}} = \begin{cases} \frac{1}{1 - p_n}S_{i,j} & \text{if } i = j \& i \notin Z_b, \\ \frac{1}{1 - p_b}S_{i,j} & \text{if } i = j \& i \in Z_b, \\ S_{i,j} & \text{if } i \neq j. \end{cases} \] (14)

And

\[ E [P]^{\text{unbiased}}_{i,j} = S_{i,j} \] (15)

These two computed matrices will then be used to learn the mapping in each layer in our proposed smSDA.
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 Zbare 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 cyber language. 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-dimension-al vectors to represent semantics of words. The well-trained word embeddings lie in a vector space where similar words are placed close to each other. 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. 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 office 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_i w_j, we simply use an additive model to derive the corresponding embedding as follows:

\[ v(w_i w_j) = v(w_i) + v(w_j) \]  \hspace{1cm} (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 Zbare 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. For the \( r \)th feature, the corresponding

\[ F_r = \frac{\sum_{i=1}^{c} n_i \left( \mu_i - \mu \right)^2}{\sum_{i=1}^{c} n_i \sigma_i^2} \]  \hspace{1cm} (22)

where \( c \) denotes the number of classes and \( n_i \) represent the number of data in class \( i \). \( \mu \) and \( \bar{\mu} \) denote the mean of entire data and class \( i \) for the \( r \)th feature, and \( si \) 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.

3.4smSDA 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 nl posts are labeled and the corresponding vector of binary labels is y = fy1; : : : ; ynlg. The binary label 1 or 0 indicates the post is or is not a cyberbullying one. Here, nl _ n, which means the labeled posts have a small size. The bullying feature set Zhs 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 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 Zbfor layer one. The experiments in Section 4 will show that the construction of the set Zb is very simple and efficient with litter human labor. For the subsequent layers, after obtaining the output of each layer, the set Zb is updated using feature ranking with Fish score according to Eq. Based on predefined dropout probabilities for bullying features and other normal features pb and pnon the bullying feature set Zb, we compute these two expected matrices E [P] and E [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 Tmax times, where the initial value for 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 Z 2 Rd(L+1)_n following Eq. (9). It is clear that the new space has a dimension of (L + 1)d. A linear SVM is trained on the training corpus, i.e. the first nl columns in 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 denoising auto encoder (SDA), as an unsupervised representation learning method, is able to learn a robust feature space. In SDA, the feature correlation is explored by the reconstruction of corrupted data. The learned robust feature representation can then boost the training of classifier and finally improve the classification accuracy. In addition, the corruption of data in SDA actually generates artificial data to expand data size, which alleviate the small size problem of training data.

2) Forcyberbullying 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 W for each layer, considering

Each word is only correlated to a small portion of the whole vocabulary. We formulate the solution for the mapping weights W as an Iterated Ridge Regression problem, in which the semantic dropout noise distribution can be easily marginalized to ensure the efficient training of our proposed smSDA.

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.

IV. 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:

TwitterDataset: Twitter is "a real-time information net-work 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 August 31, 2011 and manually labeled. 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. Some examples of bullying traces are shown in Figure 3. To preprocess these tweets, a tokenizer is applied without any stemming or stop word 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 the statistics of this dataset can be found in Table 1.

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 get multiple windows. 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 texts for these data, as XML files, have been kindly provided by Kontosta et al. 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.  
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

<table>
<thead>
<tr>
<th>TABLE 1. Statistical Properties of the two datasets.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Feature No.</td>
</tr>
<tr>
<td>Sample No.</td>
</tr>
<tr>
<td>Bullying Instances</td>
</tr>
</tbody>
</table>

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. @USERAME hahaha hopefully! Believers 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!!

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

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.

4.2 Experimental Setup

Here, we experimentally evaluate our smSDA on two cyber-bullying 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 Nicholson's performance of the Joker

Figure 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 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 Following the original setting, we scale the bullying features by a factor of 2.
- * mSDA : marginalized stacked denoising autoencoder.
- * smSDA and smSDAu: semantic-enhanced

Margin alizedm denoising auto encoder 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 _ for document topic multinomial and hyperparameter _ for word topic multinomial to 1 and 0.01, respectively. For mSDA5, the noise intensity is set to 0.5 and the number of layers for Tweets and MySpace datasets are both set to 2. Here, t number of layers is only set to be a
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 su datasets constructed from Twitter and MySpace datasets, respectively. Since BWM does not require training documents, its results over the whole 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 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 existences of bullying words are defined as rules for classification. It shows that only an elaborated utilization of such bullying words in stead of a simple one can help cyberbullying detection. We also compare our methods with two state-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. We also compare our methods with two state-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 hyper parameter 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. 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.

![Figure 8. Classification Accuracies and F1 Scores of All Compared Methods on Twitter Datasets](image)

We compare the performances of smSDA and smSDAu, which adopt biased semantic dropout noise and unbiased semantic dropout noise, respectively. The results have shown that smSDAu 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 xi that are used to compute mapping weights are the same, which cannot 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.

![Figure 9. Data Base F1 and F2](image)

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 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.

| Table 3. Term Reconstruction on Twitter data sets. 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). |
|-----------------|----------------|----------------|----------------|----------------|
| BullyingWords   | reconstructed words | mSDA | mSDA u | mSDA | mSDA u |
| hitch           | 0          | USER           | USER          | USER           | USER          |
| friend          | 1          | McGill         | McGill        | McGill         | McGill        |
| all             | 2          | McGill         | McGill        | McGill         | McGill        |
| gets            | 3          | McGill         | McGill        | McGill         | McGill        |
| fucking         | 4          | McGill         | McGill        | McGill         | McGill        |
| friend          | 5          | McGill         | McGill        | McGill         | McGill        |
| off             | 6          | McGill         | McGill        | McGill         | McGill        |
| pissed          | 7          | McGill         | McGill        | McGill         | McGill        |
| shit            | 8          | McGill         | McGill        | McGill         | McGill        |
| gets in         | 9          | McGill         | McGill        | McGill         | McGill        |
| friend          | 10         | McGill         | McGill        | McGill         | McGill        |
| off             | 11         | McGill         | McGill        | McGill         | McGill        |
| if              | 12         | McGill         | McGill        | McGill         | McGill        |

![Table 3](image)
V. 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 auto encoder as a specialized representation learning model for cyberbullying detection. In addition, word embeddings have been used to automatically expand and refine a bullying word list 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 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 Form SDA

<table>
<thead>
<tr>
<th>Bullying Words Reconstructed</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>bitch @USER @USER shut HTTPLINK friend fuck up tell shut fucking because off friend pissed off shit gets of shit some abuse big this shit with shit lollol big learned representation by considering word order in messages.</td>
<td></td>
</tr>
</tbody>
</table>

VI. REFERENCES


