IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

NCAIT 2017

JSS Academy of Technical Education
Vol. 4, Special Issue 8, May 2017

Cyberbullying Detection Based on Semantic Enhanced Marginalized Denoising Auto-Encoder

Sahana.B.R[1], Prof. Jagadisha. N[2]

Student, Dept of ISE, East West Institute of Technology, Bangalore, India[1]

Professor, Dept of ISE, East West Institute of Technology, Bangalore, India[2]

Abstract: As a side effect of increasingly popular social media, cyberbullying has emerged as a serious problem afflicting children, adolescents and young adults. Machine learning techniques make automatic detection of bullying messages in social media possible, and this could help to construct a healthy and safe social media environment. In this meaningful research area, one critical issue is robust and discriminative numerical representation learning of text messages. In this paper, a new representation learning method is introduced to tackle this problem. The method named Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) is developed via semantic extension of the popular deep learning model stacked denoisingautoencoder. 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. The proposed method is able to exploit the hidden feature structure of bullying information and learn a robust and discriminative representation of text.

Key words: Cyberbullying Detection, Text Mining, Representation Learning, Stacked DenoisingAutoencoders, Word Embedding.

I. INTRODUCTION

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. Cyberbullying on social media can take place anywhere at any time. For bullies, they are free to hurt their peers' feelings because they do not need to face someone and can hide behind the Internet. For victims, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or social media.

Previous works on computational studies have shown that natural language processing and machine learning are powerful tools to study bullying.

Cyberbullying detection can be named as a supervised learning problem. A classifier is first trained on 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. This paper focuses on text-based cyberbullying detection.

In cyberbullying detection, the numerical representation for Internet messages should be robust and discriminative. With the knowledge of one deep learning method named stacked denoising auto encoder (SDA). In this paper investigates a new text representation model based on SDA: marginalized stacked denoisingautoencoders (mSDA), which adopts linear instead in order to learn more robust representations

We utilize semantic information to expand mSDA and develop Semantic-enhanced Marginalized Stacked DenoisingAutoencoders (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, i.e. correlation, between bullying and normal words.Our proposed Semantic-enhanced Marginalized Stacked DenoisingAutoencoder 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.

* Semantic information is incorporated into the reconstruction process. In our framework, high-quality semantic information, i.e., bullying words, can be extracted automatically through word embeddings. Finally, these specialized modifications makes it easier for bullying detection.

* Comprehensive experiments on real-data sets have verified the performance of the proposed model.

II. RELATED WORK

Text representation and automatic cyberbullying detection both play a vital role in this work.

➢ 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.

IEEE TRANSACTIONS ON AFFECTIVE COMPUTING 3 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 Latent Dirichlet Allocation are also proposed. Topic models try to define the generation process of each word occurred in a document. Similar to the approaches afore mentioned, our proposed approach takes the BoW representation as the input. However, this 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 learned representation can be more robust. Third, specific to cyberbullying detection, our method employs the semantic information, including bullying words and sparsity constraint imposed on mapping matrix in each layer and this will in turn produce more discriminative representation.

➢ Cyberbullying Detection

With the increasing popularity of social media in recent years, cyberbullying has emerged as a serious problem afflicting children and young adults. Previous studies of cyberbullying focused on extensive surveys and its psychological effects on victims, and were mainly conducted by social scientists and psychologists. Although researchers have said that the machine learning is gaining increased popularity in recent years. Several research areas including topic detection and affective analysis are closely related to cyberbullying detection. 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 post. Dinakar et.al used Linear Discriminative Analysis to learn label specific features and combine them with BoW features to train a classifier. In addition, they need to construct a bully space knowledge base to boost the performance of natural language processing method. 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.

III. PROPOSED WORK

SEMANTIC-ENHANCED MARGINALIZED STACKED DENOISING AUTO-ENCODER

We first introduce notations used in this paper. Let D = {w1,...,wd}be the dictionary covering all the words existing in the text corpus. We represent each message using a BoW vector x ∈Rd. Then, the whole corpus can be denoted as a matrix:
X = [x1,...,xn] ∈Rd×n, where n is the number of available posts. We next briefly review the marginalized stacked denoising auto-encoder and present our proposed Semanticenhanced Marginalized Stacked Denoising Auto-Encoder.

➢ Marginalized Stacked Denoising Auto-encoder .

The basic idea behind denoising auto-encoder is to reconstruct the original input from a corrupted one ˜ with the goal of obtaining robust representation.

➢ Semantic Enhancement for mSDA

The advantage of corrupting the original input in mSDA can be explained by feature co-occurrence statistics. The co-occurrence information is able to derive a robust feature representation under an unsupervised learning framework, and this also motivates other state-of-the-art text feature learning method such as Latent Semantic Analysis and topic models. As shown in Figure 1. (a), a denoisingautoencoder is trained to reconstruct these removed features values from the rest uncorrupted one. It is shown that the learned representation is robust and can be regarded as a high level concept feature since the correlation information is invariant to domain-specific vocabularies.

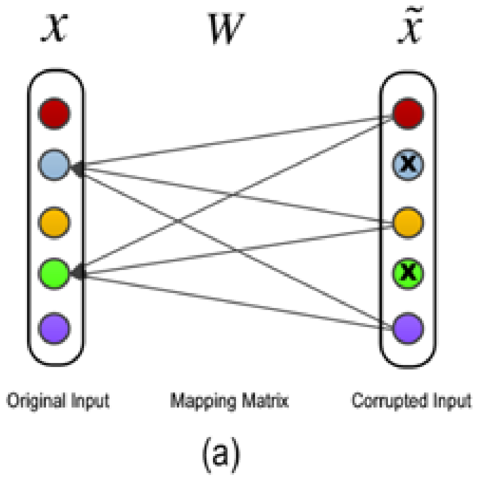


Fig. 1(a)

In Fig 1.(a) and 1(b) The cross symbol denotes that its corresponding feature is corrupted.

➢ 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 foul languages. Cyberbullying words can be explored 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. This kind of dropoutnoise 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. The proposedsmSDA can dealwith the problem learning a robust feature representation, which is a high level concept representation. 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.

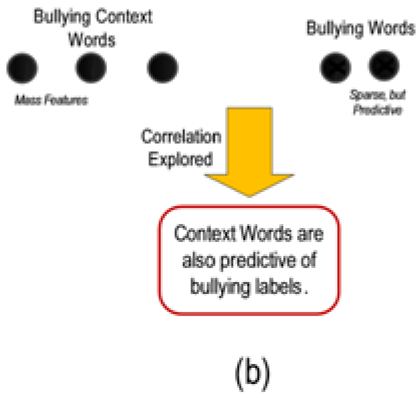


Fig. 1 (b)

➢ Construction of Bullying Feature Set

As analyzed above, the bullying features play an important role and should be chosen properly. In the following, the steps for constructing bullying feature set are given, in which the first layer and the other layers are addressed separately. For the first layer, expert knowledge and word embeddings are used. For the other layers, discriminative feature selection is conducted. Layer One: firstly, we build a list of words with negative affective, including swear words and dirty words. Then, we compare the word list with the BoW features of our own corpus, and regard the intersections as bullying features. However, it is possible that expert knowledge is limited and does not reflect the current usage and style of 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-dimensional vectors to represent semantics of words. The well-trained word embeddings lie in a vector space where similar words are placed close to each other. In addition, since the word embeddings adopted here are trained in a large scale corpus from Twitter, the similarity captured by word embeddings can represent the specific language pattern, For example, the embedding of the misspelled word fck is close to the embedding of fuck so that the word fck can be automatically extracted based on word embeddings.

The proposed work has the following steps:

- ❖ Firstly, the user has to register by signing up as in an online social network, by entering his/her userid and password. After registrations the users can login with their authentication.fig 2(a)
- ❖ Secondly, once after the registration is done he /she can login by providing credential details. Fig 2(b)
- ❖ Where after the existing users can send messages to privately and publicly, options are built. Users can also share post with others. The user can able to search the other user profiles and public posts. In this module users can also accept and send friend requests. Fig 2(c)
- ❖ If in case, the user posts information containing bullying words or foul language the post gets blocked and will not be displayed. Fig 2(d)

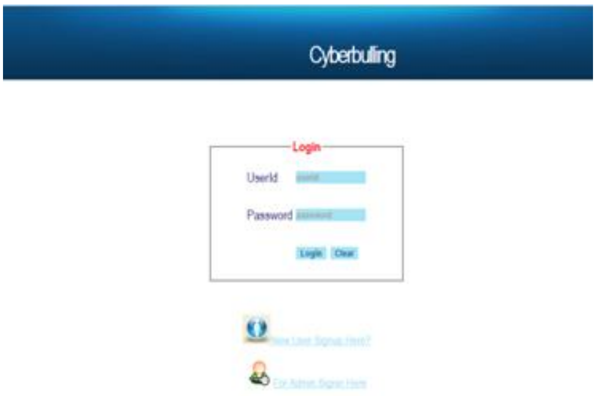
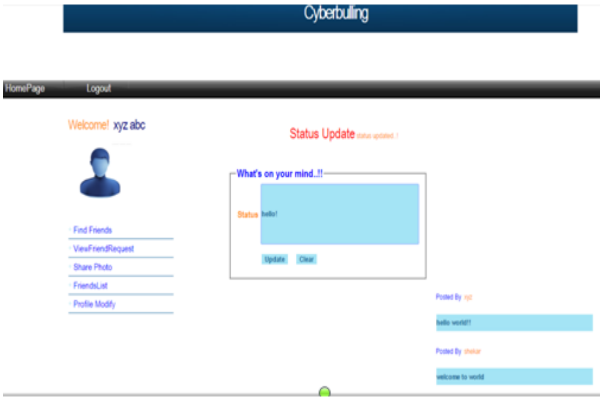


Fig.2(a) Example of a snapshot of a login page.



hello!

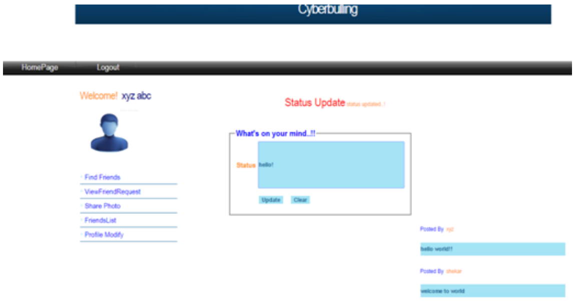
Posted By xyz

hello world!!

Posted By shekar

welcome to world

Fig.2(b) Example of a snapshot of a sign-up page

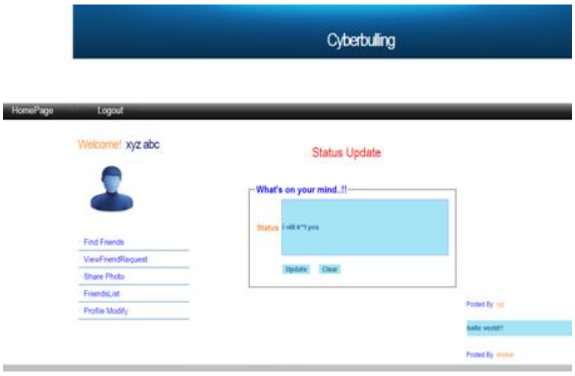


hello!

hello world!!

welcome to world

Fig.2(c)Example of a snapshot of status updated



i will k**l you

hello world!!

Fig. 2(d) Example of a snapshot of status update denied.

IV. MERITS OF sMSDA

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

- Stacked densoingautoencoder (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.

- For cyberbullying problem, we design semantic dropout noise to emphasize bullying features in the new feature space, and the yielded new representation is thus more discriminative for cyberbullying detection.

- 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.

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 bullying word lists that are 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.