Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 8, August 2017, pg.89 – 94

Detection of Text based Cyberbullying using Semantic Enhanced Marginalized Denoising Autoencoder Learning Model

Veeramallu Naga Srinivas[1], Veerendra Bethimeedi[2]

[1]M.Tech Student (CSE), 14JQ1D5806, Kakinada Institute of Technology and Science, Divili, East Godavari, Andhra Pradesh, INDIA
[2]HOD, Asst Professor, M.Tech CSE department, Kakinada Institute of Technology and Science, Divili, East Godavari, Andhra Pradesh, INDIA

[1] nagasrinivasv17@gmail.com, [2] veeru506@gmail.com

Abstract— As more people turn to the Internet for school, work, and social use, so too do more people turn to the Internet to take out their frustrations and aggression. One form of cyber aggression has been gaining the attention of both researchers and the public in recent years: cyberbullying. Cyberbullying is typically defined as aggression that is intentionally and repeatedly carried out in an electronic context (e.g., e-mail, blogs, instant messages, text messages) against a person who cannot easily defend him- or herself. Many researchers have noted that cyberbullying is occurring at widespread rates among youth and adults, with some studies showing nearly 75% of school-age children experiencing this form of aggression at least once in the last year. The experience of cyberbullying has been linked with a host of negative outcomes for both individuals and organizations (e.g., schools), including anxiety, depression, substance abuse, difficulty sleeping, increased physical symptoms, decreased performance in school, absenteeism and truancy, dropping out of school. To deal with these problems, In this paper, we investigate one deep learning method named stacked denoising autoencoder (SDA). We develop a 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. 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.

Keywords— Cyberbullying, Autoencoder, Denoising, Bullying, Feature set.

I. INTRODUCTION

There is no question that the Internet and related technologies have revolutionized the way that our world operates [1]. The popularity of the Internet among school-age children and adolescents has become apparent to most, as nearly all youth between 12 and 17 use the Internet, and 68% of school pupils use the Internet at school. Further, youth spend an average

of about 17 hours per week on the Internet, with some spending more than 40 hours per week online. Although most youth spend time communicating with their friends online, including forging new online friendships, online interpersonal interactions can be particularly valuable for those who experience anxiety in face-to-face interactions.

Although the Internet has certainly provided many benefits, it may be responsible for a host of negative outcomes as well. Among youth who venture online, almost a third report being contacted by someone they did not know through the Internet, and many report that this contact made them feel uncomfortable. Other research has found a link between duration of Internet use and psychiatric symptoms, with those reporting more Internet use also experiencing more depression, obsessive compulsion, and. The question of directionality with respect to this association clearly bears scrutiny, but the association appears robust. Furthermore, the Internet has provided some with an avenue through which to commit various counterproductive behaviors, such as cyber-hacking (i.e., using the Internet to gain access to information or resources illegally), cyber-stalking (i.e., using the Internet to spy on or watch another person), and various forms of cyber aggression including cyberbullying (i.e., using the Internet to harm another person [2]. Additionally, some individuals may develop "pathological technology use" . PTU refers to obsessive and addictive behaviors in response to technological media, such as the Internet or gaming, that resemble behaviors characteristic of addictions to alcohol or drugs.

Certain features of online communications, including reproducibility, lack of emotional reactivity, perceived uncontrollability, relative permanence, and 24/7 accessibility, make it more likely for online misbehavior to occur. With regard to reproducibility, the core issue is that a person can easily copy all of his or her friends on a message or forward gossip to his or her entire address book. This reproducibility may make it easy for deviant individuals to harm others and to repeat the harm over and over again with the click of a button. Communications over the Internet also feature a lack of emotional reactivity. When people communicate face-to-face, they provide many verbal and nonverbal cues about how they are feeling. For example, frowns or eyebrow raises are common nonverbal cues that are used when a conversation has upset the receiver. If such a cue is accurately perceived by the message sender, the sender might soften his or her message or seek clarifying feedback. In an online context, communicators do not have this instant emotional reactivity, and they might more easily offend others in their communications.

As noted earlier, most researchers agree that cyberbullying involves the use of electronic communication technologies to bully others. However, as will be seen, assessments of the prevalence of cyberbullying have proven difficult because there is a lack of consensus regarding the more specific parameters by which cyberbullying should be defined [3]. Conceptualizing cyberbullying is compounded by the fact that cyberbullying can take so many different forms and occur through so many different venues. [4] has created a taxonomy of types of cyberbullying that includes flaming (i.e., an online fight), harassment (i.e., repetitive, offensive messages sent to a target), outing and trickery (i.e., soliciting personal information from someone and then electronically sharing that information with others without the individual's consent), exclusion (i.e., blocking an individual from buddy lists), impersonation (i.e., posing as the victim and electronically communicating negative or inappropriate information with others as if it were coming from the victim), cyber-stalking (i.e., using electronic communication to stalk another person by sending repetitive threatening

communications), and sexting (i.e., distributing nude pictures of another individual without that person's consent). The media through which cyberbullying can occur are equally diverse, including instant messaging, e-mail, text messages, web pages, chat rooms, social networking sites, digital images, and online games.

This paper addresses the text-based cyberbullying detection problem, where robust and discriminative representations of messages are critical for an effective detection system. By designing semantic dropout noise and enforcing sparsity, we have developed semantic-enhanced marginalized denoising autoencoder as a specialized representation learning model for cyberbullying detection. In addition, word embeddings have been used to automatically expand and refine bullying word lists that is initialized by domain knowledge.

II. RELATED WORK

Cyber bullying is emerging as a serious social problem, especially among teenagers. Cyber bullying is defined as "the use of information technology to harm or harass other people in a deliberate, repeated, and hostile manner". With the advent of social media networks such as Twitter and Facebook, it has become more prevalent. Thus, automatic detection of cyber bullying posts is becoming an increasingly important area of research among social media researchers. Previous researches on cyber bullying detection have mostly used text-based methods and employed contextual and sentiment features to improve the text mining system. For instance, Reynolds et al. [5] used the number, density and the value of foul words as features to determine the cyber bullying messages. Similarly, Dinakar et al. [6] found that building individual topic-sensitive classifiers improves the detection of cyber bullying messages. Recently, Dadvar et al. [7] also presented an improved model using the user-based features (i.e: the history of the user's activities). All these works are based on text mining. Recently, Nahar et al. [8] built the cyber bullying network graph model and used a ranking method to identify the most active cyber bullying predators and victims. However, the cyber-bullying detection aspect of this work was still purely text-driven.

The accuracy of text-based cyber bullying detection methods still remains limited. In this paper, we advance the state of the art by adopting a more holistic approach. Our main goal is to explore the value of social information in detecting cyber bullying above and beyond the signals available in the textual content of messages. We believe that since bullying is a social problem, information about the social context surrounding the messages might provide vital clues for their detection. Using a corpus of Twitter messages, our approach identifies both social and textual features and creates a composite model for detecting cyber bullying. The obtained results suggest that the social signals are useful for detecting cyber bullying, and that using multiple channels of information (text plus social features) results in higher detection performance.

III.PROPOSED WORK

The proposed work is divided into following modules:

- Marginalized Stacked Denoising Auto-encoder
- Semantic improvement for mSDA
- Construction of Bullying Feature Set
- smSDA for Cyberbullying Detection

Marginalized Stacked Denoising Auto-encoder: It can planned a changed version of Stacked Denoising Auto-encoder that employs a linear instead of a nonlinear projection therefore on acquire a closed-form answer . The basic idea behind denoising auto-encoder is to reconstruct the first input from a corrupted one ~x1,…., ~xn with the goal of obtaining sturdy illustration. Marginalized Denoising Auto-encoder: In this model, denoising auto-encoder attempts to reconstruct original information mistreatment the corrupted information via a linear projection. Illustration of Motivations behind smSDA.

Semantic Improvement for mSDA: The advantage of corrupting the original input in mSDA may be explained by feature co-occurrence statistics. The co-occurrence information is in a position to derive a strong feature illustration below associate degree unsupervised learning framework, and this also motivates different progressive text feature learning ways such as Latent linguistics Analysis and topic models. A denoising auto encoder is trained to reconstruct these removed features values from the rest uncorrupted ones. Thus, the learned mapping matrix W is able to capture correlation between these removed options and different options. The major modifications include linguistics droupout noise and distributed mapping constraints. However, a direct use of those bullying options might not come throughs} good performance as a result of these words solely account for a little portion of the complete vocabulary and these vulgar words area unit just one quite discriminative features for bullying.

Construction of Bullying Feature Set: The bullying features play associate degree vital role and will be chosen properly. In the following, the steps for constructing bullying feature set Zb are given, in which the primary layer and therefore the different layers area unit self-addressed individually. For the first layer, expert data and word embeddings area unit used. For the other layers, discriminative feature selection is conducted. Layer One: firstly, we build a list of words with negative affection, including swear words and dirty words. Then, we compare the word list with the BoW options of our own corpus, and regard the intersections as bullying features. and does not replicate the present usage and elegance 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 linguistics of words. The well-trained word embeddings lie in a vector space wherever similar words area unit placed near one another. In addition, the cosine similarity between word embeddings is in a position to quantify the linguistics similarity between words. Considering the Internet messages are our interested corpus, we utilize a well-trained word2vec model on a large-scale twitter corpus containing four hundred million tweets. A visualization of some word embeddings when spatial property reduction (PCA). It is observed that curse words kind distinct clusters, which area unit additionally remote from traditional words. Even insulting words are set at totally different regions due to different word usages and insulting expressions. In addition, since the word embeddings adopted here are trained in a massive 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 in order that the word fck may be mechanically extracted supported word embeddings.

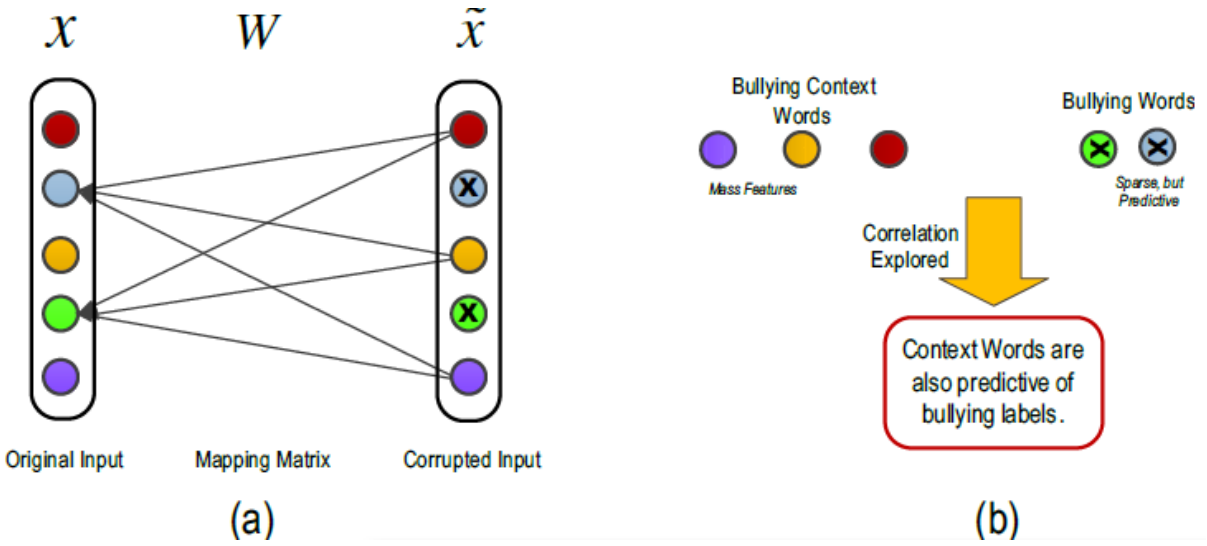

Figure 1

smSDA for Cyberbullying Detection 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. In Figure 1(a), the cross symbol denotes that its corresponding feature is corrupted, i.e., turned off. Some important merits of our proposed approach are summarized as follows:

1) Most cyberbullying detection methods rely on the BoW model. Due to the sparsity problems of both data and features, the classifier may not be trained very well. Stacked densoing autoencoder (SDA), as an unsupervised representation learning method, is able to learn a robust feature space. In SDA, the feature correlation is explored by the reconstruction of corrupted data. The learned robust feature representation can then boost the training of classifier and finally improve the classification accuracy. In addition, the corruption of data in SDA actually generates artificial data to expand data size, which alleviate the small size problem of training data.

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

3) The sparsity constraint is injected into the solution of mapping matrix 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. CONCLUSIONS

Cyber Bullying, which often has a deeply negative impact on the victim, has grown as a serious issue among adolescents. To understand the phenomenon of cyber bullying, experts in social science have focused on personality, social relationships and psychological factors involving both the bully and the victim. Recently computer science researchers have also come up with automated methods to identify cyber bullying messages by identifying bullying-related keywords in cyber conversations. However, the accuracy of these textual feature based methods remains limited. Our proposed Semantic-Enhanced Marginalized Denoising Auto-Encoder can deal with the problem by 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.

REFERENCES

[1] Menesini, E., Calussi, P., & Nocentini, A. (2012). Cyberbullying and traditional bullying: Unique, additive, and synergistic effects on psycho-logical health symptoms. In Q. Li, D. Cross, & P. K. Smith (Eds.), Cyberbullying in the global playground: Research on international perspectives (pp. 245–262). Malden, MA: Blackwell.

[2] Holfeld, B., & Grabe, M. (2012). An examination of the history, prevalence, characteristics, and reporting of cyberbullying in the United States. In Q. Li, D. Cross, & P. K. Smith (Eds.), Cyberbullying in the global playground: Research from international perspectives (pp. 117– 142). Malden, MA: Blackwell.

[3] Paul, S., Smith, P. K., & Blumberg, H. H. (2012). Revisiting cyberbullying in schools using the quality circle approach. School Psychology International, 33, 492–504. doi:10.1177/0143034312445243

[4] Williams, K. R., & Guerra, N. G. (2007). Prevalence and predictors of Internet bullying. Journal of Adolescent Health, 41(6, Suppl.), S14–S21. doi:10.1016/j.jadohealth.2007.08.018

[5] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on, volume 2, pages 241–244. IEEE, 2011.

[6] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In The Social Mobile Web, 2011.

[7] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. Improving cyberbullying detection with user context. In Advances in Information Retrieval, pages 693–696. Springer, 2013.

[8] V. Nahar, X. Li, and C. Pang. An effective approach for cyberbullying detection. Communications in Information Science and Management Engineering, 3(5):238–247, 2013.

