International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 25 Issue 4 – APRIL 2018.

CYBERBULLYING DETECTION BASED ON SEMANTIC-ENHANCED MARGINALIZED STACKED DENOISING AUTO-ENCODER

Sparsha R[#1],Sushma M Balgi[#2], Amulya Anand[#3], Sushma K B [#4] and B B Neelkantappa[*5]

[#] *Student, Computer Science and Engineering, Malnad College of Engineering, Hassan, India.*

[*] *Associate Professor, Department of Computer Science and Engineering, Malnad College of Engineering, Hassan, India.*

***Abstract*— The rapid growth of social networking is supplementing the progression of cyberbullying activities. Most of the individuals involved in these activities belong to the younger generations, especially teenagers, who in the worst scenario are at more risk of suicidal attempts. Cyberbullying is the process of using the Internet, cell phones or other devices to send or post text or images intended to hurt or embarrass another person. Through machine learning techniques, we can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyberbullying content. . Here, we introduce a new machine learning method to deal with this problem. Our method named Semantic-Enhanced Marginalized Stacked Denoising Auto-Encoder (smSDA) is developed via semantic extension of the popular deep learning model. The smSDA method detects the hidden attributes of the bullying information. Our approach is experimented on two public cyberbullying corpora i.e, Twitter and MySpace. The outcome of our proposed method is better than the other text representation learning methods.**

***Index Terms*— Cyberbullying, Machine learning, smSDA, Semantic extension.**

I. INTRODUCTION

Social Media, is defined as "a group of Internet based applications that build on the ideological and technological base and that allow the creation and interchange of user-generated information". 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 mass via digital communication methods such as sending texts and posting comments against a victim. 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 victim, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or sociable media. One way to tackle the cyberbullying trouble is to automatically detect and promptly report bullying messages so that proper measures can be taken to prevent possible tragedies. Cyberbullying sensing can be formulated as a supervised learning problem. In the United States, approximately 43 percent of teenagers were ever bullied on social media. The same as traditional bullying, cyberbullying has negative, insidious and sweeping encroachment on children. The outcomes for victim under cyberbullying may even be tragic such as the occurrence of self-injurious behaviour or suicide.

Three kind of information including textual matter, user demography, and social network place 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 erudition of text content.

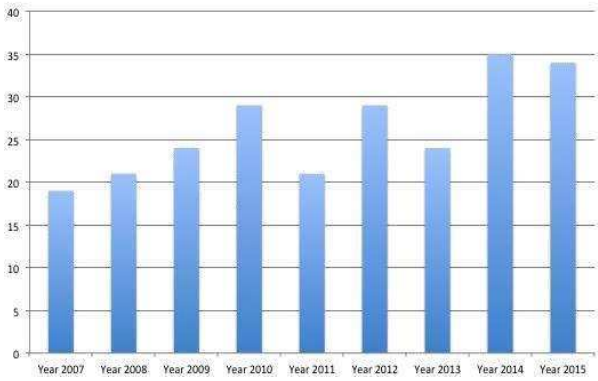

Figure 1.1 Percentage of teens who were cyberbullied across America.

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 popular text matter representation models, which are 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 signification behind text

units.

In cyberbullying detection, the numerical representation for net content should be robust and discriminative. Since messages on social media are often very short and contain a band of informal language and misspelling , robust representations for these messages are required to reduce their ambiguity. The main aim of this study is to develop methods that can learn robust and discriminative representations to address the problems in cyberbullying detection. We make use semantic information to expand mSDA and develop smSDA (Semantic-Enhanced Marginalized Stacked Denoising Auto-Encoder). The semantic data consists of bullying words.

An instinctive extraction of bullying words based on word embedding is proposed so that the involved human effort can be reduced. During training of smSDA, we try to rebuild the bullying features from other usual words by learning the latent structure, i.e., correlation, between bullying and usual words. The perception behind this idea is that some bullying messages do not include bullying words. The correlational statistics information discovered by smSDA helps to reconstruct bullying features from usual words, and this in turn facilitates detection of bullying information without containing bullying words. For example, there is a strong correlation between bullying word kill and normal word off since they often occur together. If bullying message do not contain such obvious bullying characteristic, such as kill is often misspelled as k*ll, the correlation may help to reconstruct the bullying characteristic from normal ones so that the bullying message can be detected.

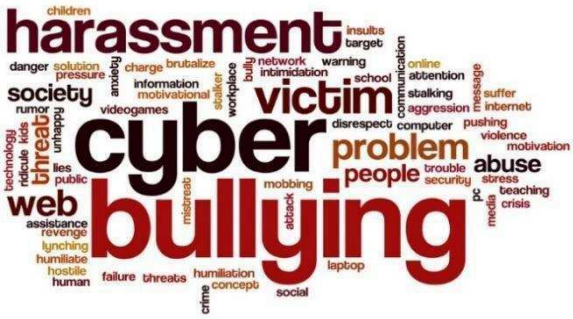

Figure 1.2 Cyberbullying

II. LITERATURE SURVEY

The success of machine learning algorithmic program generally depends on information representation, and we hypothesize that this is because different representations can entangle and hide more or less the different explanatory factors of variation behind the information [1].

The notion of Social Media is top of the agenda for many business executive today. Based on this definition, we then provide a classification of Sociable Media which groups application program currently subsumed under the generalized term into more specific groups by characteristic: collaborative projects, blogs,content communities, societal networking sites and virtual social world [2].

It provides a acute review of the existing cyberbullying research. The general aggression model is proposed as a useful theoretical model from which to understand this phenomenon. Additionally, results from a meta-analytic review are presented to highlight the sizing of the relationships between cyberbullying and traditional bullying, as well as relationships between cyberbullying and other expressive behavioural and psychological variables [3].

It provides the overall detecting problem into subtle topics, lending itself into text edition classification sub-problems. We experiment with a corpus of 4500 YouTube remark, applying a range of binary and multiclass classifier. We break through that binary classifiers for individual recording label outperform multiclass classifiers [4].

Cyberintimidation detection can be expressed as a supervised learning problem. A classifier is first trained on a cyberbullying principal labeled by humans, and the learned classifier is then used to recognize a bullying message [5].

III. CYBERBULLYING DETECTION BASED ON SMSDA

Most cyberbullying detection procedures depend on the BoW model. Due to the sparsity problems of both data and features, the classifier may not be trained very well. Stacked denoising autoencoder, as an unsubstantiated representation learning method, is able to learn a robust feature space.

For cyberbullying problem, we design semantic dropout noise to emphasize bullying lineament in the new feature space, and the yielded new representation is thus more discriminative for cyberbullying sensing. Based on word embedding, bullying features can be extracted automatically.

In addition, the possible limitation of proficient knowledge can be alleviated by the use of word embedding.

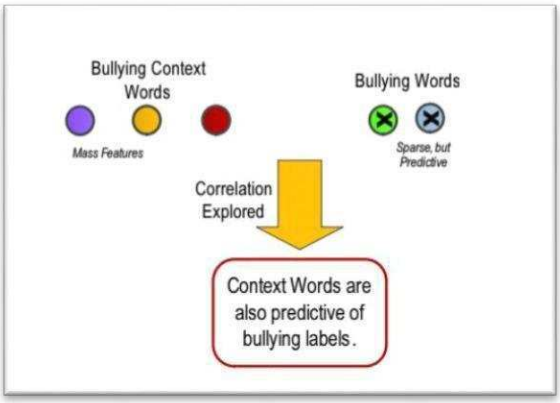

Figure 3.1 System Design Overview Modules

Modules :

- Construction of OSN System
- Bullying Feature Set Construction
- Cyberbullying Detection.

- Semantic-Enhanced Marginalized Stacked Denoising Auto-Encoder.

A. Construction of OSN System :

In the first part, we develop the Online Social Networking (OSN) system. We construct the system with the characteristics of Online Social Networking. Here, this module is used for new the user registrations. The user can login with their respective username and password for the authentication purpose.

The existing users can send messages privately and publicly. Users can also share post with others. The user can able to find the other user profiles and public posts. In this module users can also accept and send friend requests. With all the basic characteristics of Online Social Networking System modules is build up in the initial module, to prove and evaluate our system features.

B. Bullying Feature Set Construction:

The bullying structures play a vital role and should be selected properly. In the following, the steps for creating bullying feature set are given, in which each layers are addressed separately.

For the first layer, proficient information and text embedding are used. For the remaining layers, the discriminative feature selection is conducted.

In this module, we construct a list of words with negative effects, 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.

Finally, the constructed bullying features are used to train the first level in our proposed smSDA. It includes two parts: one is the master copy insulting seed based on domain knowledge and the other is the extended bullying words via word embedding.

C. Cyberbullying Detection :

In this part, we propose the Semantic-Enhanced Marginalized Stacked Denoising Auto-encoder (smSDA). Here, we describe how to leveraging it for cyberbullying detection. smSDA provides vigorous and discriminative demonstrations. The learned numerical representations can then be provided into our system.

In the new space, due to the captured feature correlation statistics and semantic information, even trained in a small size of , grooming principal is able to achieve a good performance on testing papers.

Based on text embedding, bullying features can be extracted automatically. In addition, the possible limitation of proficient information can be alleviated by the use of word embedding.

D. Semantic-Enhanced Marginalized Stacked Denoising Auto-Encoder:

An automatic lineage of intimidation news based on phrase embedding is proposed so that the involved human labour can be reduced. During training of smSDA, we attempt to rebuild bullying features from other convention words by discovering the latent structure, i.e. connection, between bullying and usual words. The perception behind this mind is that some bullying messages do not contain bullying words.

The correlativity data discovered by smSDA helps to reconstruct intimidation features and this in turn facilitates detection of intimidation messages without containing bullying words. For example, there is a strong relation between bullying word kill and normal word off since they often occur together. If bullying messages do not contain such obvious bullying description, such as kill is often misspelled as k*ll, the correlation may help to rebuild the bullying description 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 information set, including breeding data size, which helps improve the data sparsity problem.

IV. PERFORMANCE COMPARISON WITH EXISTING METHODS

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

The demerits of the existing system are:

- The first and also critical step is the numerical illustration learning for text edition subject matter.

- Secondly, cyberbullying is hard to describe and judge from a third view due to its intrinsic indistinctness.

- Thirdly, due to security of internet users and privacy issues, only a small share of messages are left hand on the Internet, and most bullying position are deleted.

The proposed system has the following merits:

- Our proposed Semantic-Enhanced Marginalized Stacked Denoising Autoencoder is able to learn robust characteristics from BoW methods in an efficient and effective way. These robust lineament are learned by reconstructing original input from corrupted (i.e., missing) ones. The new feature infinite

can improve the performance of cyberbullying detection even with a small labelled training corpus.

Dataset	Measures	BWM	BoW	sBoW	LSA	LDA	mSDA	smSDA
Twitter	Accuracies	69.3	82.6	82.7	81.6	81.1	84.1	84.9
	Scores	16.1	68.1	62.3	65.8	66.1	70.0	71.9
MySpace	Accuracies	34.2	80.1	80.1	77.7	77.8	87.8	89.7
	Scores	36.4	41.2	42.5	45.0	43.1	76.1	77.6

TABLE 4.1 TABLE OF ACCURACIES FOR COMPARED METHODS ON TWITTER AND MYSPACE DATASETS

- Semantic entropy is incorporated into the reconstruction process via the designing of semantic dropout noise and imposing sparsity constraints on chromosome mapping matrix . In our framework, high-quality semantic data, i.e., bullying words, can be extracted automatically through word embedding.

- Finally, these particular changes make the new characteristic space more discriminative and this in turn facilitates bullying detection.

The equivalence of our proposed smSDA method with six benchmark approaches on Twitter and MySpace datasets and the average results, for these two datasets, on sorting accuracies are shown in Table 4 .1 and the corresponding graph is given in figure 4.1.

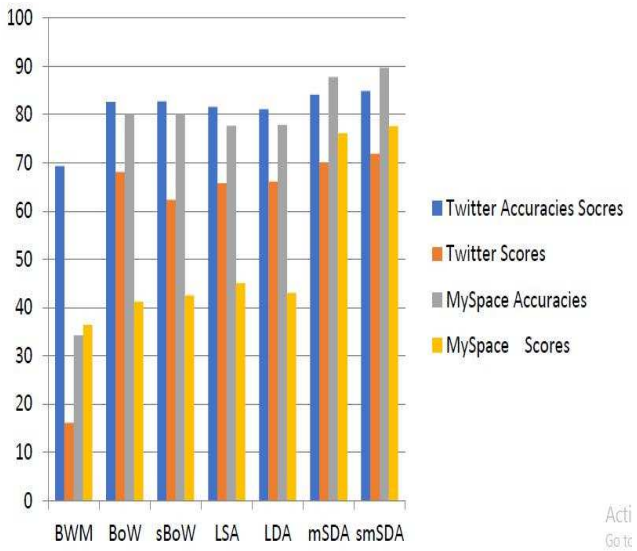

Figure 4.1: Graph of Accuracies for Compared Methods on Twitter and MySpace.

V. EXPECTED RESULTS

The first notice is that semantic Bow structure (sBow) performs slightly better than BoW. Based on BoW, sBoW just arbitrarily scale the bullying features by a element 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 move towards smSDA, gains a significant performance improvement compared to sBoW. This is because bullying characteristic only accounting for a small share of all features used. It is difficult to learn robust features for small training data by intensifying each bullying features range.

Our approach target to find the correlation between usual bullying features by reconstructing corrupted data so as to issue robust features. In addition, Intimidation Word Matching, as a simple and intuitive method of using semantic information, gives the worst performance.

In BWM, the existence of bullying words is defined as rules for sorting. It appears that only an elaborated usage of such bullying words instead of a simple one can help cyberbullying sensing.

VI. CONCLUSION

The project addresses the text-based cyberbullying tracking problem, where robust and discriminative structure of information are critical for an effective detection system. By design semantic dropout noise and enforcing sparsity, we have developed smSDA as a specialized representation learning model for cyberbullying detection. In addition, word embedding have been used to automatically expand and refine bullying word lists that are initialized by field knowledge.

The performance of our approach has been experimentally verified through two cyberbullying corpora from sociable medias: Twitter and MySpace. As a next step we are planning to further improve the robustness of the learned representation by considering order of words in the message.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[2] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media, Bus. horizons, vol.53, no. 1, pp. 59- 68, 201
[3] R. M. Kowalski, G. W. Giumentti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and metaanalysis of cyberbullying research among youth," Physchol. Bulletin, vol. 140, pp. 1073-1137, 2014.

[4] K. Dinakar, R. Reichart, and H. Lieberman, "Modelling the detection of textual cyberbullying", presented at the Social Mobile Web, Barcelona, Spain, 2011.
[5] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyberbullying detection using social and textual analysis", in Proc. 3rd Int. Workshop Socially-Aware Multimedia, 2014, pp. 3-6.
[6] T. H. Dat and C. Guan, "Feature selection based on Fisher ratio and mutual information analysis for robust brain computer interface", in Proc. IEEE Int. Conf. Acoust, Speech, Signal Process, 2007, vol. 1, pp. I-337-I-340.
[7] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification. New York, NY, USA: Wiley, 2012.
[8] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, p. 27-1–27-27, 2011.
[9] J. Sui, "Understanding and fighting bullying with machine learning," Ph.D. dissertation, The Univ. of Wisconsin-Madison, WI, USA, 2015.
[10] J. Bayzick, A. Kontostathis, and L. Edwards, "Detecting the presence of cyberbullying using computer software," in Proc. Int. Conf. Web Sci., Koblenz, Germany, Jun. 2011, pp. 1–2.